December 2016

Assessing Model-Data Fit for Compensatory and Non-Compensatory Multidimensional Item Response Models Using Vuong and Clarke Statistics

Leanne Freeman
University of Wisconsin-Milwaukee

ASSESSING MODEL-DATA FIT FOR COMPENSATORY AND NON-COMPENSATORY MULTIDIMENSIONAL ITEM RESPONSE MODELS USING VUONG AND CLARKE STATISTICS

by

Leanne L. Freeman

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Educational Psychology

at

University of Wisconsin-Milwaukee

December 2016

ABSTRACT

ASSESSING MODEL-DATA FIT FOR COMPENSATORY AND NON-COMPENSATORY MULTIDIMENSIONAL ITEM RESPONSE MODELS USING VUONG AND CLARKE STATISTICS

by

Leanne L. Freeman

The University of Wisconsin-Milwaukee, 2016
Under the Supervision of Professor Bo Zhang

The primary importance of the study was that no statistics were able to provide probabilistic statements about model-data fit between the non-nested compensatory and non-compensatory MIRT models. Secondarily, the Vuong and Clarke statistics have been utilized prolifically in economics and political science and have great potential to contribute to educational measurement. Finally, application of the Vuong and Clarke statistics will not only reduce the damage of misspecified MIRT models but will also promote the use of less known models, in this case, the non-compensatory models.

The purpose of the study was to investigate whether the Vuong and Clarke statistics can be used to detect model-data fit between compensatory and non-compensatory MIRT models. The effectiveness of the statistics was evaluated through simulated Type I error and power studies. The Type I error studies compared the true and estimated compensatory models. The power studies compared the true non-compensatory model with the estimated compensatory model. The controlling factors included test structure, sample size, test length, and correlation between person traits. Overall, the statistics produced very large values which resulted in, on average, extremely high rejection rates for all conditions, if the assumed sampling distributions

of the two fit statistics were used. In other words, the nominal Type I error rates were not observed. Consequently, alternate processes were employed to assess statistical power: The Receiver Operating Characteristic (ROC) curves were used to assess the discrimination ability of the statistics and a Monte Carlo resampling technique was used to assess power rates.

Results of both analyses clearly indicated the value of the Vuong and Clarke statistics in detecting model-data fit for the MIRT models. The ROC curve results provided evidence of discrimination ability of both statistics under most conditions. The power analyses provided strong evidence that under most conditions the Vuong statistic, the Clarke statistic, or both statistics are able to detect misfit. There was an observable impact of the four condition factors on the performance of the fit statistics. The patterns of results included increased power with increased test length and sample size, and decreased power with increased correlation between person traits. The statistics were particularly effective with large sample size, large test length, and low correlation between trait conditions. The statistics were less effective when correlation between traits was high. There was also a difference in the performance of the statistics across test structure.

TABLE OF CONTENTS

Abstract ii

List of Figures vii

List of Tables viii

Acknowledgements ix

CHAPTER 1: INTRODUCTION 1

CHAPTER 2: LITERATURE REVIEW 6

Multidimensional Item Response Theory (MIRT) 6

Test Structure 14

Model Estimation 17

Model Data Fit 18

CHAPTER 3: METHODS 26

Research Design 26

Test Characteristics 28

Data Simulation 30

Model Estimation 32

Computation of Model Fit Statistics 33

CHAPTER 4: RESULTS 34

Null Conditions 34

Power Conditions 39

ROC Results 44

Power Results 47

CHAPTER 5: DISCUSSION 54

Limitations and Further Research 60

Conclusions 62

REFERENCES 63

CURRICULUM VITAE 71

LIST OF FIGURES

Figure 1. Compensatory 2PL model ICS with parameters $a_1 = 1.5$, $a_2 = 0.5$, $d = 0$ 12

Figure 2. Non-compensatory 2PL model ICS with parameters $a_1 = 1.5$, $a_2 = 0.5$, $d_1 = 0$, $d_2 = 0$. 12

LIST OF TABLES

Table 1. Type 1 Error and Power Study Breakdown....26

Table 2. Null Conditions, Descriptive Statistics for Vuong Statistic....36

Table 3. Null Conditions, Comparing Test Structure for Vuong Statistic....37

Table 4. Null Conditions, Descriptive Statistics for Clarke Statistic....38

Table 5. Null Conditions, Comparing Test Structure for Clarke Statistic....39

Table 6. Power Conditions, Descriptive Statistics for the Vuong Statistic....41

Table 7. Power Conditions, Comparing Test Structure for Vuong Statistic....42

Table 8. Power Conditions, Descriptive Statistics for Clarke Statistic....43

Table 9. Power Conditions, Comparing Test Structure for Clarke Statistic....44

Table 10. AUC Values for Vuong Statistic....46

Table 11. AUC Values for Clarke Statistic....47

Table 12. Power Cutoff Values for Approximately Simple Structure for Vuong and Clarke Statistics....51

Table 13. Power Cutoff Values with Complex Structure for Vuong and Clarke Statistics....51

Table 14. Power for Vuong Statistic....52

Table 15. Power for Clarke Statistic....52

Table 16. Power....53

ACKNOWLEDGEMENTS

There are many people I'd like to thank for their support and guidance during my doctoral studies:

First, I'd like to thank my advisors. My deep gratitude to my major advisor, Dr. Bo Zhang, for his time, patience, insights, and commitment, especially when the going got tough. To my minor advisor, Dr. David Armstrong, my sincere appreciation for his enthusiasm and willingness to share his knowledge and experience. My advisors have been incredibly influential and I will forever be grateful for their involvement throughout this adventure.

Next, I'd like to thank my dissertation committee, Dr. Cindy Walker, Dr. Razia Azen, and Dr. Stephen Wester. My success in the classroom and beyond is in large part due to the knowledge they imparted and the academic experiences they afforded me. Their instruction and friendship has been, and always will be, invaluable.

I'd also like to thank my friends and family who have been so incredibly supportive during my many years of being a student. They shared in my joy over the successes, helped me laugh during the challenges, and encouraged me every step of the way. I am where I am today because of their unwavering love and support.

A very special thank you to my mother, Deanna Freeman-Gorman. There is no way to express in words my appreciation. She always believed I could achieve anything and with her support, honesty, love, and encouragement this dissertation was written.

CHAPTER 1: INTRODUCTION

Item response theory (IRT) became a viable research and test construction option in educational testing in the early 1970s. Previously, classical test theory (CTT; Croker & Algina, 1986; Lord & Novick, 1968; Traub, 1997) was favored as it was easy to use and interpret, but it had some fundamental drawbacks which IRT strove to address. As early as 1960, IRT was presented in basic model form (Reckase, 1997) with the idea that item and ability characteristics could produce probabilistic statements about responses made by an examinee to an item. The theories and models originally proposed were expanded upon (e.g., more dimensions, different outcome types; Lord & Novick,1968), but it wasn't until a decade later that computing speed and efficiency of computers was advanced enough to handle the procedures. As advancements in computer technology exploded so did the availability of computer programs, the accessibility to, and implementation of IRT (Yen & Fitzpatrick, 2006).

Like most model-based theories, IRT has a number of assumptions with the most fundamental being unidimensionality. Unidimensionality states that a test should measure only one person trait, yet the question of whether or not "a one-dimensional latent space (is) an adequate practical approximation for actual test items?" was often asked (Lord & Novick, 1968; p. 381); thus began the conceptualization of multidimensionality. A multidimensional item is one which is able to measure two or more abilities within a single item. As an extension of IRT, the models addressing multidimensional items came to be known as Multidimensional Item Response Theory (MIRT) models.

One of the advantages of MIRT models is that because the examinee can produce multiple scores, the results are more informative. For instance, a math item that assesses both geometric and arithmetic knowledge produces one score for geometric ability and another for

arithmetic. These separate scores allow for more precise assessments. Unfortunately, even though MIRT models may be used in computer adaptive testing, dimensionality assessment, test linking and equating, and proficiency classifications, just to name a few (Zhang & Stone, 2007), the actual use of them is still negligible (Walker & Beretvas, 2001). In general, current practices still subscribe to the assumption of unidimensionality, yet many argue that most items, regardless of field of use, are multidimensional in some way and that when the number of skills needed to answer any item is really dissected unidimensionality is actually quite rare (Bolt & Lall, 2003; Traub, 1997). Continued study of MIRT advances the understanding of the models and promotes their application.

Within MIRT there are two distinct types of models to be considered: the compensatory and non-compensatory. These two types are mathematically and theoretically different. The mathematical form of the compensatory MIRT model is attributed to Reckase (1985) and uses summation of the parameter values to calculate probability of a correct answer. On the other hand, the non-compensatory model uses multiplication to calculate the probability of a correct answer. The mathematical form of the non-compensatory MIRT model is attributed to Sympson (1978; Ackerman, 1992, 1994). The equations will be presented in detail.

Theoretically the models differ on how they answer the question of how an examinee uses multiple traits or dimensions to answer an item. The compensatory model has its theoretical foundation in that multiple traits work together and that strength of one trait can compensate for weakness of the other. The non-compensatory model has its theoretical foundation in the idea of trait independence. If one trait is weak, the probability of answering the item correctly will be low, regardless of the strength of the other trait(s). Until recently, computer programs mainly supported compensatory models because the non-compensatory model is more computationally

complex and parameter estimation more difficult. Because of this and the comparative general ease of estimation and interpretation, the compensatory model has been the default choice in most MIRT applications. A problem is that, as with unidimensional models not always being the best fitted, the compensatory model may not always be the best fitted model. If examinees are answering an item using abilities progressively rather than simultaneously or if knowledge on one trait doesn't help with lack of knowledge in another, it would mean that the traits are actually non-compensatory instead of compensatory. This would result in model misspecification, which can have serious consequences. One way to avoid misspecification is to conduct model-data fit analyses when there is a viable alternate model option.

The most common model-data fit comparisons are performed between nested models. A model is nested in another when they are linearly related and one can be reduced to the other by imposing linear restrictions. Models are nested only if one model is a subset of the other and they are theoretically similar (Clarke, 2001; Osteen 2010). In IRT, either unidimensional or multidimensional, there are 3 popular dichotomous outcome models frequently used. These models have only two outcomes which are often coded as yes (1) or no (0). These include the 1-parameter (1PL), the 2-parameter (2PL), and 3-parameter (3PL) logistic models. The 1PL is nested in the 2PL, and the 2PL is nested in the 3PL. Common ways to compare these models, regardless of whether the models are in the uni- or multi-dimensional family (Ackerman et al., 2003), is through the chi-square test of the likelihood difference between the full model (e.g., the model with more parameters) and the restricted model.

However, traditional model-data fit statistics like those mentioned cannot be used when comparing non-nested models (Clarke, 2001). This is because non-nested models, by definition, are not linearly related, one is not a reduction of the other, and they are mathematically and/or

theoretically different (Clarke, 1998). As will be presented in detail, the compensatory and non-compensatory models are mathematically and theoretically different which identifies them as non-nested. Also, with non-nested models, there is no natural hypothesis and the distributions may belong to different families (Clarke, 1998; Pesaran & Deaton, 1978; Pesaran & Ulloa, 2007).

Since the traditional model-data fit statistics cannot be used to compare non-nested models, researchers use informal and ad hoc decision criteria like Akaike's information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Clarke, 1998; Schwarz, 1978). For model-data fit comparison between compensatory and non-compensatory models, these types of indices are the only option at his time. The issue with these indices is that there is no associated significance and the best model is the one with the relatively smaller value. That being said, there are two promising statistics: The Vuong statistic (1989) and the Clarke statistic (2003). Both have a long history and application in political science and economics (Genius & Strazzera, 2001; Hall & Pelletier, 2011; Hong & Preston, 2005) and even though the Vuong statistic has been in use for far longer than the Clarke statistic, the overall research suggests that both tests show good properties in a variety of conditions (Clarke, 2003; Clarke & Signorino, 2010; Pesaran & Ulloa, 2007) and may be applied in a wide range of fields (e.g., health insurance; Czado et al., 2014). The current study provided information on the properties and applicability of the Vuong and Clarke statistics for use in yet another field of study, educational testing.

The main goal of the current study was to provide evidence whether the Clarke and Vuong statistics are viable options for detecting model-data fit between compensatory and non-compensatory MIRT models. To achieve this, the two statistics were studied in various testing

conditions through Monte Carlo simulations. Their effectiveness was evaluated by both Type I error and statistical power.

CHAPTER 2: LITERATURE REVIEW

Multidimensional Item Response Theory (MIRT)

Item response theory (IRT) and multidimensional item response theory (MIRT) models relate latent variables or traits to the probability of a correct response on items. Traits may be defined by aptitude, achievement, or personality variables (Ackerman, 1994; Ackerman et al., 2003; Hambleton et al., 1991). The goal is to describe the interaction between persons and items by assigning a score to an examinee that is representative of the examinee's level of ability on the trait (IRT) or traits (MIRT) being measured by the item. An example of a unidimensional item is an equation style problem that is mathematics with no text. To answer this type of item one might only need to use an ability called "algebraic symbol manipulation" (Ackerman, 1994, p. 256). An example of a multidimensional item is a word problem that assesses both language comprehension *and* algebraic knowledge. One of the assumptions of IRT modeling is unidimensionality which states that only one trait is being assessed by the item. Of importance is that, as can be seen from the examples, some items may assess more than one ability. When test items measure more than one trait, either because it is established a priori or because the dimensional structure is unclear, the unidimensional assumption will be violated and MIRT models become applicable (Beguin & Glas, 2001).

The argument has been made that most items in cognitive, educational, and psychological tests are more often than not multidimensional and that dimensionality should always be verified and not assumed (Ackerman, 1994; Ackerman et al., 2003; Traub 1997). This being the case, there are potentially many times when unidimensional IRT models are being used incorrectly (Ackerman, 1992). The improper use and application of a model may lead to serious validity issues. This in one of the many reasons there has been a push for more research into MIRT

models (Hambleton et al., 1991, Segall, 2001). Another reason for this push is to close the gap inherent between the theoretical developments of MIRT and their applications (Hulin et al., 1985). MIRT methods are important because they provide information about the nature of the goal construct assessed by the test, aid in decisions about what type of scores or sub-scores to report, allow for application of model-data fit options beyond limited-information methods, shed light on the item-ability interaction, provide support for the test's construct validity, they have the flexibility to be used in a variety of applications (e.g., test development, diagnostics, differential item functioning, etc.), and the ability to accurately represent and interpret examinee perceptions and outcomes (Ackerman et al., 2003; Allen & Wilson, 2006; Bolt, 2001).

The importance of dimensionality can be illustrated by studies by Allen and Wilson (2006) and Vollema and Hoijtink (2012). Allen and Wilson performed a health study exploring dimensionality. The goal was to see if the construct of self-regulation is better measured on a single dimension or on two dimensions (e.g., type of regulation and motivation). The tool used for analysis was an assessment of self-regulation that has six sub-tests. For unidimensionality the scores were based on a composite approach (sum of the scores across all subtests) and a consecutive approach (scores within each subtest summed and each subtest treated in separate analyses). The multidimensional model was a version of the compensatory MIRT Rauch model (see equation 4). The results of the study indicated that self-regulation was best measured by the multidimensional structure. The authors contended that by using the MIRT model the advantages (versus the other two unidimensional approaches) included less error, separate subscores for both dimensions, and increased reliability of the estimates. Vollema and Hoijtink (2012) also explored dimensionality by analyzing the psychological measure Schizotypal Personality Questionnaire (SPQ) to ascertain if the dichotomous items were assessing 2 or 3 dimensions. The self-report

measure was completed by outpatient and inpatient adult psychiatric patients. As with the previous study, versions of the multidimensional Rauch model were utilized. The results indicated that schizotopy clearly has 3 dimensions which were defined as positive schizotypy, negative schizotypy, and disorganization. The authors found the 3-dimensional structure to be stable across psychotic and non-psychotic individuals, the dimensions clear and useful, and supported by previous research.

As stated earlier, MIRT models are extensions of unidimensional IRT models. A detailed discussion of the dichotomous outcome IRT models, from most complex to least, will be presented now. Dichotomous models are those which have and outcome of either 1 (e.g., correct answer) or 0 (e.g., incorrect answer). A discussion of the MIRT models will follow. The most complex of the unidimensional IRT models is the three-parameter logistic model (3PL). The mathematical form of the 3PL model is

$$P\left(U_{ij}=1\middle|\theta_j,a_i,b_i,c_i\right)=c_i+(1-c_i)\frac{1}{1+e^{-1.7a_i(\theta_j-b_i)}} \quad (1)$$

where $P(U_{ij}=1|\theta_j,a_i,b_i,c_i)$ is the probability that an examinee j with ability θ will answer item i correctly, U_{ij} is item response, a_i is the item discrimination parameter or slope for item i, b_i is the item difficulty/location parameter, and c_i is the guessing parameter. There are 3-item characteristic parameters (e.g., a_i, b_i, c_i) and 1-person parameter (e.g., θ_j) for this model. Without guessing, the three-parameter model will reduce to the two-parameter logistic model (2PL). Its mathematical form is

$$P(U_{ij}=1|\theta_j,a_i,b_i)=\frac{1}{1+e^{-1.7a_i(\theta_j-b_i)}} \quad (2)$$

where all parameters are defined above. The final models in this group are the one-parameter logistic model (1PL) and the Rasch model. The mathematical form of the 1PL is

$$P(u_{ij}=1|\theta_j,\alpha,b_i)=\frac{1}{1+e^{-1.7\alpha(\theta_j-b_i)}} \tag{3}$$

where the *a*'s are equal for all items. The mathematical form of the Rasch model is

$$P(u_{ij}=1|\theta_j,b_i)=\frac{1}{1+e^{-(\theta_j-b_i)}} \tag{4}$$

where all *a*'s are equal to 1.

As with unidimensional IRT, there are 4 dichotomous MIRT models which include a 3PL (M3PL), a 2PL (M2PL), a 1PL (M1PL), and a Rasch model for both compensatory and non-compensatory options (to be discussed). Each model has its own distinct quality. For example, among these three models, the M1PL model is most restrictive. It poses a high demand on test data as all items have the same discrimination parameter. The M3PL model, on the other hand, is complex due to the additional guessing parameter. The guessing parameter increases the estimation difficulty, which in turn requires more information from the test, such as large sample size. The M2PL gives the best picture of how the model fit statistics can be applied, the right amount of variability in the parameters but enough stability, and has been established as more appropriate for many testing conditions (Zhang, 2012). Therefore, the M2PL models will be explored in this study.

The first of the two MIRT models is the compensatory 2PL model (C2PL). It is defined as

$$P(U_{ij}=1|\boldsymbol{\theta}_j,\mathbf{a}_i,d_i)=\frac{1}{1+e^{-1.7(\mathbf{a}_i\boldsymbol{\theta}'_j+d_i)}} \tag{5}$$

where $\boldsymbol{\theta}_j$ is a 1 x *m* parameter vector of ability values, $\mathbf{a}_i$ is a 1 x *m* parameter vector of discrimination values (vector extensions from the IRT model), *m* is the number of dimensions, and d_i is a scalar intercept parameter. If d_i is redefined as $d_i = -ab$, then the model becomes

$$P(U_{ij} = 1 | \theta_j, a_i, b_i) = \frac{1}{1+e^{-1.7a_i(\theta_j - b_i)}} \quad (6)$$

Under the compensatory model the abilities work conjunctively to increase the probability of a correct response to an item, so higher ability on one trait compensates for lower ability on the other. For example, an examinee is asked to read a passage on a current event, a local election, and answer a question about it. This item assesses two abilities, reading comprehension and knowledge of current events. If the examinee is knowledgeable about the election, then that compensates for lower reading ability. On the flip side, if an examinee is an excellent reader then their reading skills would compensate for lack of election knowledge.

The second of the two MIRT models is the non-compensatory 2PL model (N2PL). It is defined as

$$P(U_{ij} = 1 | \boldsymbol{\theta}_j, \mathbf{a}_i, \mathbf{b}_i) = \prod_{\ell=1}^{m} \frac{1}{1+e^{-1.7a_{i\ell}(\theta_{j\ell} - b_{i\ell})}} \quad (7)$$

where $\mathbf{b_i}$ is a vector of difficulty parameters, ℓ is the dimension or trait of interest, and all other parameters are defined above. The form of the N2PL can be simplified to $k = \prod_{\ell=1}^{m} p_\ell$ where k is the probability of a correct response (a constant) and p_ℓ is the product of the 2PL terms. It can be further simplified to $k = p_1 p_2$. Under the non-compensatory model, traits do not compensate each other. This means that an examinee needs a high level on all traits in order to have a high chance of answering an item correctly and "the probability of correct response for an item that follows this model can never be greater than the probability for the component with the lowest probability" (Reckase, 2009, p. 98). An example of a non-compensatory item is the traditional word problem where a text passage is presented and a mathematical outcome is required. Such a test assesses two abilities, reading comprehension and math computation knowledge. If an

examinee has excellent reading ability but low math computation ability, the examinee will be able to read the text but not be able to solve the problem. The other scenario is that the examinee has high math computation ability but low reading ability. In this case, the examinee will not be able to solve the problem because they are unable to ascertain the question being asked (Sympson, 1978).

Even though the two models look very similar (after redefining the *d* parameter in the compensatory model), the difference is in the operation. For example, the parameters for an item are set to $a_1 = 1.5$, $a_2 = 0.5$, $b = 0$, $\theta_1 = -1$, $\theta_2 = 2$ for both models. The probability of a correct answer based on the compensatory is 0.299. To find the probability of a correct answer for the non-compensatory model (assume $b_1=b_2=0$), the components are multiplied to arrive at a probability of .06. This makes clear the difference in the models. This examinee has low ability on dimension 1 with a higher ability on dimension 2, so by using the compensatory model the higher ability makes the probability of a correct answer greater than if using the non-compensatory model. But with the non-compensatory model, the low ability on dimension 1 forces a low probability of a correct answer.

From this example, it is clear that an important question in how an item is correctly answered is whether the interaction of abilities is compensatory or not. The compensatory model is more holistic in its hypothesis and examinees use resources collectively to come to an answer. On the other hand, the non-compensatory model has more of a separatist view where success is achieved by using different skills and knowledge to answer different parts of the item. There are theoretical implications which provide information on what processes are being applied and what strategies are being used (Hartig & Hohler, 2009).

The differences between the models are well illustrated by the graphical representations of item characteristic surfaces (ICS). In general, the visual representation of the models must take into account the item characteristics in a multidimensional space and because these models have 2 dimensions, the visualization of the relationship between item and person must support both dimensions simultaneously. Examples of an ICS for theoretical compensatory and non-compensatory models are shown in Figures 1 and 2, respectively. The ICS for the compensatory model (Figure 1) shows the compensatory nature of the relationship. In looking at the graph one can see that even if an examinee gets a lower score on θ_2 (e.g., 0), they can still have a high probability of answering the item correct is the score on θ_1 is greater than 1.5. Looking at the non-compensatory model (Figure 2), one can see that even with a high score on the first ability, an examinee with a score of zero on the second ability will not answer this item correctly. Where the compensatory model has an entire area of high probability, the non-compensatory has a point of high probability and the scores on both abilities must be large.

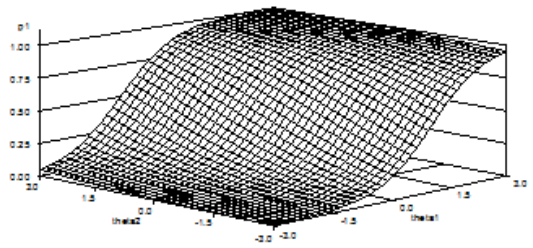

Figure 1. Compensatory 2PL model ICS with parameters $a_1 = 1.5$, $a_2 = 0.5$, $d = 0$

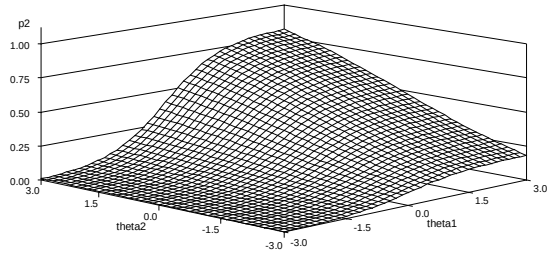

Figure 2. Non-compensatory 2PL model ICS with parameters $a_1 = 1.5$, $a_2 = 0.5$, $d_1 = 0$, $d_2 = 0$.

Compensatory models have been researched and applied more in practice than non-compensatory models (Ackerman, 1994; Ackerman et al., 2003; Babcock, 2011; Reckase, 1985). There are a few possible reasons: the theory of compensation has been readily accepted or assumed (Svetina, 2012), the models are more user-friendly with several computer programs available for parameter estimation, and they have been used with solid and consistent results. However, there are situations when the data is actually better fitted to the non-compensatory model. When this occurs and compensatory models are still used, misfit results which in turn causes serious repercussions including lack of validity, assumption violations, and problems with estimation (McKinley & Mills, 1985; Stone, 2000). Until recently it has been difficult to examine and use the non-compensatory models as there have been no programs available for calculation and estimation. Things have evolved in that there are now options for running non-compensatory models, yet there have been very few studies focused on the models.

That being said, the interest in and amount of research has increased in recent years. A study performed by Bolt and Lall (2003) evaluated parameter recovery for the compensatory and non-compensatory 2PL models. The factors were sample size (1,000 and 3,000), test length (25 and 50 items), and correlation between traits (0, .3, and .6). The results indicated that the parameters for both models were successfully, although somewhat inconsistently, estimated but, in general, the compensatory model had fewer errors, better estimated when correlation between abilities was high, and smaller samples were needed.

Babcock (2011) explored the estimation of the non-compensatory 2PL model using the Bolt and Lall (2003) study as a road map. The primary independent variables were correlation between traits (.25, .50, .75) and sample size (1,000, 2,000, and 4,000). Multidimensional conditions had a total of 50 items and there were unidimensional components that varied in

length. The results indicated that extremely large sample sizes (e.g., N = 4,000), low correlation between traits, and more than two unidimensional items per dimension were necessary for accurate estimation. Even with large sample sizes, accurate estimations were difficult to obtain with high correlation between traits.

In yet another study (Spray et al., 1990) the goal was to determine if there was a difference in the models at the item and test levels. In other words, are the more similar than not and are they actually measuring different processes? The study factors were a test length of 20-items, sample size N = 2,000, and correlation between traits of .0, .25, .50, and .75. At the test level, the results indicated that at higher correlations the models were almost indistinguishable. At the item level, it was found that the amount of error increased with increased correlation between traits. Also, the difference in models decreased with increased correlation between traits but, in general, a distinct difference between the models was observed.

The results of these studies illustrate the differences between the compensatory and non-compensatory models and the importance in continuing to study them because (a) validity issues that may arise if the assumptions of compensation are incorrect, (b) the debate between whether items should be considered compensatory or not and to what degree continues (Ackerman, 1994), and (c) there is so little evidence to support one theory or the other and under what conditions they are most applicable. One reason there are few studies may be that even though there are programs that run both types of models there is a lack of validated non-compensatory estimation programs.

Test Structure

Within the multidimensional framework model structure describes the relationship between the dimensions. For the M2PL model, the structure describes how much of the

variability of an item is determined by dimension 1 and how much by dimension 2. Three structure types that are often encountered in practice are simple structure, approximately simple structure, and complex structures (Zhang & Stone, 2007).

With simple structure, an item measures only one dimension while different items measure different dimensions. With approximately simple structure, each item measures a dominant dimension and one or more secondary dimensions. Stout et al. (1996) defined items with approximate simple structure as those that "can be partitioned into item clusters that are each relatively dimensionally homogeneous and that are dimensionally distinct from each other" (p. 331). An example of a test that uses approximately simple structure is the Law School Admission Test (LSAT). The LSAT has three item types including logical reasoning (LR), analytical reasoning (AR), and reading comprehension (RC). Stout et al. (1996) analyzed the LSAT and found it was (a) multidimensional and (a) that the LR section assessed the two dimensions of additional information (AI) and hidden assumptions (HA) with a clear approximately simple structure.

For a test with complex structure the primary dimension is not as clearly defined and both (or all) abilities are used substantially to answer an item. The contribution of each dimension is subtler and diffused which makes in more difficult to detect model-data misfit as compared with approximately simple structure. Finch (2011) investigated 3PL compensatory MIRT model estimations for approximately simple and complex structures. The results indicated that the standard error increased with the more complex structure, with non-normal distribution of latent traits, with larger correlations between traits, and with smaller sample sizes.

Test structure in the multidimensional space is by calculating angular distance which can be calculated by

$$\cos \alpha_{i\ell} = \frac{a_{i\ell}}{\sqrt{\sum_{k=1}^{m} a_{ik}^2}} \tag{8}$$

The difference in the structures is the maximum angular difference with the dominant dimensions having the greater value. θ_1 is the dominant trait when a_1 has the larger value while θ_2 is the dominant trait when a_2 the larger value. For tests with approximately simple structure, the dimensions can be defined by angles 0 - 15° and 75° - 90°. For tests with complex structure, the dimensions can be defined by angles 0 - 15° and 45°- 90°. This makes the delineation of the traits more diffuse. As an example, if $a_1 = .7$, value of a_2 can take on a range of values which defines if as having approximately simple or complex structure. With a smaller value of $a_2 = .2$ the structure is approximately simple. With a larger value of $a_2 = .5$, θ_1 is still the dominant dimension, but the values are closer which defines it as having complex structure.

The impact of structure has been a subject of study. In one study, Zhang and Stone (2008) investigated the applicability of the $s - \chi^2$ statistic (Orlando & Thissen, 2000) in evaluating MIRT model-data fit. Type I error and power studies were performed using the approximately simple and complex structures. Approximately simple structure was defined as having dimension 1 angles of 0-20° and dimension 2 angles 70° to 90°. Complex was defined of having dimension 1 angles of 0-20° and dimension 2 angles 45° to 90°. It was found that with Type I error the complex structure the distributions were more difficult to approximate as compared with approximately simple structure while the results of the power studies were similar to those found with unidimensional IRT comparisons.

One of the issues with IRT and MIRT models is that of indeterminacy when parameters of the models can have multiple values that produce the same probability. By specifying the angles and restricting the range indeterminacy is addressed.

Model Estimation

Marginal maximum likelihood estimation (MMLE) is the most popular item parameter estimation method (de Ayala, 2009). The basic premise is that the ability is integrated out in the estimation of the item parameters to provide clean item estimations. MMLE has three basic steps: (a) The first step is to choose item parameter starting values, if approximations are known. The closer the values are to the real item parameter values, the faster the convergence will be. (b) The second step is to compute likelihoods, but to do so the individual thetas from the equation need to be extracted and replaced with a distribution. The person parameters are considered random effects and because the calibration sample is assumed to come from a random sample from a population, the item parameter estimations can be freed from the person parameter estimations. To approximate the continuous population distribution, a discrete distribution made of rectangles is used. In general, more rectangles lead to better approximations (deAyala, 2009). One can either divide the entire distribution into rectangles or use adaptive quadrature where only the likely section is chosen for integration. (c) The next step is the integration. For integration, for each rectangle the likelihood of each person's response pattern is weighted by that rectangle's probability of being observed. Then, across all the rectangles, the weighted likelihoods are summed (e.g., integrated). The likelihood is then maximized through an iterative process. If convergence is not attained, then the process repeats using a different starting value. If convergence is attained, the parameter values are very close to those that will provide the maximum likelihood and the ability estimates can be made.

The MMLE procedure is similar for uni- and multi-dimensional IRT item estimation. The difference is that instead of using the likelihood of a single item score to make estimations like with unidimensional IRT, MIRT estimation is based on the likelihood of a string of item scores.

That likelihood is used to calculate the probability of observing that particular item score string in a population of examinees. The philosophy behind maximum likelihood is the same for both IRT and MIRT because the estimation is still based on the notion of "the maximum of a surface is the place where the tangent plane to the surface has a slope of zero in all directions [and] the point where the tangent plane touches the surface is the maximum of the surface" (Reckase, 2009, p. 151).

MMLE only produces item parameter estimates so the ability estimates must be made using the procedures of maximum likelihood estimate (MLE) or the Bayesian methods of maximum a posteriori (MAP) or expected a posteriori (EAP). The Bayesian methods (EAP and MAP) utilize the data provided (the item parameters) to develop the posterior distribution which is then sampled through an iterative process. An advantage of the Bayesian approach is that it can be used to get location estimates for all theta response patterns including perfect and zero scores. MLE is a popular estimation approach, but it is unable to produce finite estimates when an examinee answers all items either correct or incorrect. The Bayes options are able to do this and they can produce a theta estimate for all examinees. Versus MAP, there are two main reasons that EAP is more efficient. First, EAP uses quadrature points for estimation while MAP uses a continuum; a continuous option will take more points and more time to resolve. Second, EAP uses the mean statistic for estimation while MAP uses the mode. It has been found that using the mean makes EAP less computationally demanding during the iterative process (da Ayala, 2009).

Model Data Fit

It is important to determine the best available model for the data (Hong & Preston, 2005) because when a model does not fit the data there can be violations in model assumptions or

issues with the estimation procedures. Most importantly, lack of model-data fit can negatively impact the validity of how the estimates are used (McKinley & Mills, 1985; Stone, 2000). Lack of fit can occur for many reasons including assumptions of the model not being met or problems with the estimation process. There are different procedures to assess different types of model fit in IRT (e.g., item fit, person fit, overall model fit; McKinley & Mills, 1985). In the current study the interest is in overall model fit and specifically in model-data fit for the compensatory and the non-compensatory 2PL MIRT models.

To determine the best model for the data the appropriate model-data fit statistic must be used. One of the deciding factors is the relationship between the models (e.g., nested or non-nested). Mathematically, models are nested only if two models are linearly related and one can be reduced to the other by imposing linear restrictions. In other words, if one model is a subset of the other (Clarke, 2001; Osteen 2010). Another aspect that contributes to a nested relationship is that of theoretical similarity. Alternately, when models are not linearly related, when the distribution of the two models is different, when the link functions are different, when the predictors are not hierarchical, and/or when there is a theoretical difference, the models are non-nested (Czado et al., 2014).

An example of nested models is the IRT 2PL and Rasch models. The 2PL model is defined as

$$P(U_{ij}=1|\theta_j,a_i,b_i)=\frac{1}{1+e^{-1.7a_i(\theta_j-b_i)}} \tag{9}$$

and the Rasch model is defined as

$$P(u_{ij}=1|\theta_j,b_i)=\frac{1}{1+e^{-(\theta_j-b_i)}} \tag{10}$$

The linear transformation there is a=1 for all the items.

An example of non-nested models is the compensatory and non-compensatory 2PL MIRT models. The compensatory 2PL model is defined as

$$P(U_{ij} = 1|\theta_j, a_i, b_i) = \frac{1}{1+e^{-1.7a_i(\theta_j - b_i)}} \quad (11)$$

and the non-compensatory 2PL model is defined as

$$P(U_{ij} = 1|\boldsymbol{\theta}_j, \mathbf{a}_i, \mathbf{b}_i) = \prod_{\ell=1}^{m} \frac{1}{1+e^{-1.7a_{i\ell}(\theta_{j\ell} - b_{i\ell})}} \quad (12)$$

with all parameters defined above. There exists no linear function that may transform one MIRT model to the other.

As discussed previously, the traditional model-data fit processes used for nested models cannot be used to detect model-data fit between non-nested models, so the options for model-data fit analyses are limited to comparison indexes. Two of the best known are the Bayesian information criteria (BIC; Schwarz 1978) and Akaike's information criteria (AIC; Akaike 1974).

Even though BIC has *information criteria* in the name, it is actually not an information criterion at all as it does not estimate information lost. BIC does not assume that a true model exists that represents reality, is applicable with regression models fit by maximum likelihood estimates, and may be used to compare nested or non-nested models. The BIC index estimates the Bayes factor which is the probability that the data comes from one model over the other. If the probability that the observed data given Model 1 is greater than the probability that observed data given Model 2, then Model 1 is the preferred choice. Again, this does not mean that Model 1 is a good choice, just rather a better choice. The goal is to assess relative support for the two models (Long, 1997; Fox, 2008). The best model is the one with the smallest value.

Another popular comparison index is Akaike's information criteria (AIC; Akaike 1974). AIC is based on the Kullback-Leiber (1951; Clarke, 1998) information criteria (KLIC) which is defined as

$$\text{KLIC} = E_0[\ln h_0(Y_t \mid X_t)] - E_0[\ln f(Y_t \mid X_t; \beta_*)] \tag{13}$$

where $h_0(Y_t \mid X_t)$ is the true conditional density of Y given X or the true and unknown model, and β_* is the estimate of β when $f(Y_t \mid X_t)$ is not the true model. The best model minimizes the KLIC by maximizing $E_0[\ln f(Y_t \mid X_t; \beta_*)$. AIC minimizes information loss through the maximization of the log-likelihood. The goal is to separate information from noise or error. Each model of interest will have an associated AIC value. These values can be ranked and relative comparisons be made in selecting a best model fit for the data. The best model is the one with the smallest value and the least information lost (Clarke, 1998).

Clarke (1998) described two problems with these criteria and others like it. First, these methods do not allow a probabilistic statement to be made regarding model selection, so there is a distinction between hypothesis testing or absolute discrimination and model selection criteria or relative discrimination. Second, they will always choose a model even if they are both seriously misspecified. There are two statistics that address these issues. The first is the Vuong statistic (1989), the second is the Clarke statistic (2003).

The Vuong statistic (1989) has a classical testing framework, is normally distributed asymptotically, and may be used for nested, non-nested, or overlapping models. Since the statistic was first presented in 1989 it has been extensively used in economics, political science, as well as in other fields of research (Genius & Strazzera, 2001; Hong & Preston, 2005; Pesaran & Ulloa, 2007). As with AIC, the Vuong statistic also uses the KLIC with the best model

minimizing the KLIC (detailed above). This is done by maximizing $E_0[\ln f(Y_t \mid X_t; \theta_*)]$ so that it comes as close as possible to the true model $E_0[\ln h_0(Y_t \mid X_t)]$. The Vuong test null hypothesis is

$$H_0 : E_0\left[\ln \frac{f(Y_t \mid X_t; \beta_*)}{g(Y_t \mid Z_t; \gamma_*)}\right] = 0 \tag{14}$$

where E_0 is the expectation under the true model, β_* is the estimate of β when $f(Y_t \mid X_t)$ is not the true model, and γ^* is the estimate of γ when $g(Y_t \mid Z_t)$ is not the true model. The test determines if the average likelihood ratio is significantly different from zero (Clarke, 2003). If the null is true than the ratio of the log-likelihoods will be zero, if *Model f* is the better model then the ratio will be greater than zero, and if *Model g* is the better model then the ratio will be less than zero. Even though the true or expected value is not known, Vuong (1989) showed that when assumptions are met, the expected value may be estimated by $1/n$ times the likelihood ratio statistic (Clarke, 2001; Clarke & Signiorino, 2010; Vuong, 1989). The expected value estimation is

$$\frac{1}{n} LR_n(\hat{\beta}_n, \hat{\gamma}_n) \xrightarrow{a.s.} E_0\left[\ln \frac{f(Y_t \mid X_t; \beta_*)}{g(Y_t \mid Z_t; \gamma_*)}\right] \tag{15}$$

with the statistic being

$$\text{under } H_0: \frac{LR_n(\hat{\beta}_n, \hat{\gamma}_n)}{(\sqrt{n})\hat{\omega}_n} \xrightarrow{D} N(0,1) \tag{16}$$

where

$$LR_n(\hat{\beta}_n, \hat{\gamma}_n) \equiv L_n^f(\hat{\beta}_n) - L_n^g(\hat{\gamma}_n) \tag{17}$$

and the variance is

$$\hat{\omega}_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\left[\ln \frac{f(Y_t \mid X_t; \beta_n)}{g(Y_t \mid Z_t; \gamma_n)}\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n} \ln \frac{f(Y_t \mid X_t; \beta_n)}{g(Y_t \mid Z_t; \gamma_n)}\right]^2 \tag{18}$$

The Vuong statistic is a classical model selection test of relative discrimination where the hypothesis is that of no significant difference between the models. In model selection the rival models are compared directly by some criteria and then it chooses the best model. It is not concerned if the model performs in absolute terms, what matters is the relative strength of the rival models. Even though the Vuong is still a model selection test and will chose the one model that is closest to the true specification even if both are far away from that specification, Clarke (2003) argued that it is an improvement over AIC and BIC because it does allow a probabilistic statement to be made.

The Vuong statistic has been used, for example, to discriminate between non-nested models pertaining to international relations (Clarke, 2001). The study compared two theories, structural realism and rational deterrence theory, that attempt to explain "the escalation of great-power militarized disputes" (p. 725) using two sets of models. The first set of models were highly correlated and the Vuong statistic was unable to discriminate between them. The results were inconclusive. The second set of models were only slightly correlated. Under this condition the Vuong statistic performed well and identified the structural realist model as best fitting. In yet another studies it was found that the Vuong statistic was able to correctly specify misfit with smaller sample sizes (N = 300; Genius & Strazzera, 2001) and that the ability of the Vuong statistic to identify misfit decreases with increased correlations (Clarke, 1998).

The second statistic to compare non-nested models is the Clarke statistic (2003). The Clarke statistic was specifically introduced as a non-nested model selection tool. It is a distribution free test which looks at the differences in individual log-likelihoods from non-nested models by using an altered pared sign test (Clarke, 2007). Pared sign tests are binomial test used to test the hypothesis that the probability of being greater or less than zero is the same for any

random variable from a population. The logic of the Clarke statistic is that if one model has a better fit, then the individual log-likelihoods of that model should also have a better fit. The Clarke statistic null hypothesis is

$$H_0 : Pr_0\left[\ln\frac{f(Y_t \mid X_t;\beta_*)}{g(Y_t \mid \mathrm{Z}_t;\gamma_*)} > 0\right] = 0.5 \tag{19}$$

where E_0 is the expectation under the true model. The null hypothesis states that the individual log-likelihood ratios should be evenly distributed around zero with half being smaller and half being greater than zero. "While the Vuong test determines whether or not the average log-likelihood is statistically different from zero, the [Clarke test] determines whether or not the median log-likelihood ratio is statistically different from zero" (Clarke & Signorino, 2010. p. 377). In the hypothesis, the Pr_0 represents the median of the ratio. The test statistic is defined as

$$B = \sum_{i=1}^{n} I_{(0,+\infty)}(d_i) \tag{20}$$

where $d_i = \ln f(Y_t \mid X_t;\hat{\beta}_n) - \ln g(Y_t \mid X_t;\hat{\gamma}_n)$ and I is the indicator function. B is the number of positive differences and is from a random binomial distribution with $p = 0.5$. Basically, the model with the larger count has the better discrimination and is therefore the better model.

There are some distinct advantages to the Vuong and Clarke statistics over traditional non-nested model-fit indexes like AIC and BIC. First, as mentioned previously, they allow for probabilistic statements and produce p-values that can be compared to critical values. Second, the Vuong and Clarke tests are relatively easy to calculate (to be discussed). Third, the Vuong and the Clarke statistics have a complimentary relationship because they are distributed differently (e.g., normal and binomial) and a different basis for the comparison of true to estimated models (e.g., average log-likelihood ratio and median log-likelihood ratio). Because

the Clarke statistic was created to address some of the drawbacks of the Vuong statistic (e.g., low detection of misfit with non-normal data distributions) the statistics detect misfit under some of the same conditions and, more importantly, under differing conditions. Also, because differing sample sizes, test lengths, and correlations effect efficiency of the statistics differently, this also may impact under which conditions the Vuong and Clarke statistics are better at detecting misfit.

Previous research explored these differences. Clarke (2007) found that when the underlying log-likelihood ratios are normally distributed the Clarke statistic "is only . . . 64% as efficient as the Vuong test" and the Vuong statistic is the better detector of misfit, but when the underlying log-likelihood ratios are leptokurtic the "Vuong test is only 50% as efficient as the distribution fee test" (Clarke, 2007, p. 5) and the Clarke is better at detecting misfit. The factors of the study included sample size, distance from the null, and error. The results also indicated that with large error, the Clarke statistic had greater power rates than the Vuong statistic. And, finally, as sample size increased, the difference in power rates between the tests decreased and the instances of choosing the wrong model also decreased.

CHAPTER 3: METHODS

Research Design

Altogether four simulation studies were performed, as shown in Table 1. These studies investigated Type I error and power of the compensatory MIRT model under two structure types.

Table 1

Type 1 Error and Power Study Breakdown

	Structure	
	Approximately Simple	Complex
True Model	Alternate Model	
	Compensatory	Compensatory
Compensatory	*Type I error*	*Type I error*
Non-compensatory	*Power*	*Power*

When the estimated model is the same as the true model, any rejection of the model data fit is a Type I error. The goal of the two Type I error studies was to determine how the targeted fit indices (e.g., Clarke and Vuong model fit statistic) work with test data that fit the selected model. The evaluation criterion in these two studies was whether the nominal rate of the specified Type I error level was observed or not under the assumed sampling distributions. Out of 300 random datasets generated it was expected that 15 trials would result in rejection at α = .05. Taking into account sampling error, the 95% confidence interval of the Type I error rate will be computed by the following formula

$$p \pm Z_{.95}\sqrt{\frac{p(1-p)}{N}} \pm \frac{0.5}{N} \tag{21}$$

where p is the interested proportion, Z is the value from the normal distribution that corresponds to the confidence level, N is the sample size or the number of replications in this case, and finally

the fraction $0.5/N$ is a continuity correction for a continuous normal distribution approximation of a discrete binomial variable. The CI range turns out to be [0.02, 0.08] with $N = 300$, $\alpha = .05$.

The goal of the power studies was to present evidence whether the two statistics have the sensitivity to detect model misfit. Clearly, the higher the power, the more useful a fit statistic is. As a rule of thumb, a minimum power rate of 0.8 is acceptable. An important aspect of power is that its interpretation and application is dependent on Type I error being observed. When Type I error or the assumed distributions are not observed, power based on the assumed distribution would be meaningless. Instead, alternate procedures need to be explored to assess the power of the fit statistics, such as the Receiver Operating Characteristic (ROC) curve and Monte Carlo resampling methods.

The ROC curve is a non-parametric technique which assesses the discrimination ability of statistical tests (Mair et al., 2011; Zhang & Stone, 2007). The ROC curve is used prolifically in the medical sciences but has been recently applied in IRT as a tool for model-data fit within the Rauch family of models (Mair et al., 2001) and for examining the performance of item fit indices (Orlando & Thissen, 2003). It does not give a power rate, but instead reveals the potential of the fit statistic to achieve power. Most importantly, the method does not make assumptions on the sampling distributions so it is a good alternative when the assumed distribution is unknown. To create the ROC curve, ranked pairs of fitting and non-fitting items produced by a statistic are compared. Specifically, sensitivity (true positive rate) is plotted against 1-specificity (false positive rate) to create a curve. In the current study sensitivity was defined as power and 1-specificity defined as Type I error. That comparison is then used to calculate the probability that the statistic can correctly discriminate. The resulting area under the curve, or AUC statistic, indicated the level of discrimination and is reported as the models predictive strength (Elliot &

Woodward, 2010). An AUC of .5 denotes random or chance discrimination (useless test), AUCs from .7 to .79 denote acceptable discrimination, from .8 to .89 is good discrimination, from .9 to .99 exceptional discrimination, and AUCs of 1 refer to perfect discrimination. Sample size of 300 are sufficient to obtain over 80% probabilities in detecting various differences between distributions (Mair et al., 2011).

With the potential of the fit statistics established with adequate AUC values, power can be established using a Monte Carlo resampling technique; a technique in which many replications from a specified population with known characteristics are generated to produce sampling distributions (e.g., the generated distributions for the conditions of the current study) and then the sampling distributions are compared (e.g., null versus power). This is a "nonparametric approach to statistical inference that relies on large amounts of computation rather than on the mathematical analysis and distributional assumptions of traditional parametric inference" (Mooney, 1997, p. 570). In general, the first step is to calculate the upper and lower limit cut points for the null distribution. When α=.05, for a two-tailed test, the upper limit cut point is the 97.5% value of the distribution and lower limit cut point is the 2.5% value. The results from the power distributions are then compared to the cut limits and any values smaller than the lower cut point and any values larger than the upper cut point are rejected. The percentage of rejected values is power. In the current study, only the upper limit cut points of 97.5% were calculated because the null distributions were found to be at the left of the power distributions and no power values fell below the lower limit cut point.

Test Characteristics

The dichotomous two-parameter MIRT model was used for all studies. The four study factors investigated were test structure, correlation between traits, test length, and sample size.

Two tests structures were investigated including the approximately simple and complex structures. The simple structure was excluded due to the fact that the compensatory and non-compensatory models are equivalent under such a structure. For tests with approximately simple structure, the dimensions were defined with angular distance between 0 and 15° for the primary dimension and 75° and 90° for the second dimension. For tests with complex structure, the dimensions were defined by angles 0 to 15° for the primary and 45° to 90° for the secondary.

The correlations investigated were .0, .3, .5, and .8. These coefficients represent no, low, medium, and high correlations realistically present between dimensions (Yao, 2013). Correlations greater than .8 were not included because these extremes have been shown to be essentially uni-dimensional, so in these cases using a multidimensional model has no benefit (Yao, 2013). A correlation of zero is not common in reality but served as a baseline and mimics a simple structure.

The tests investigated were 20- and 40-items, representing medium to long test length (DeMars, 2004). Generally, as test length increases, the accuracy and consistency of the ability and item parameter estimates also increases, which in turn impacts the effectiveness of model fit statistics (de la Torre & Patz, 2002; Yao, 2013). Research also indicates that IRT models with fewer items have excessive Type I error rates (DeMars, 2004). So, with other factors being constant, it was expected that the 40-item conditions would have better Type I error rates and higher power.

The sample sizes investigated were 500 and 1,000. Because estimation of MIRT models requires large sample sizes these fairly large sample sizes were included. As with test length, increased sample size impacts the accuracy of the parameter estimates and in turn the efficacy of the model fit statistics (Yao, 2013). DeMars (2007) and Yao (2013) found that sample sizes of

1,000 yield good estimates. With other factors being constant, it was expected that the 1,000-subject conditions would produce larger power rates.

To summarize, the factors investigated were:

1. Test structure (approximately simple and complex)
2. Correlation between dimensions (.0, .3, .5, and .8)
3. Test length (20 and 40)
4. Sample size (500 and 1,000)

which yielded a total of 16 experimental conditions for each of the four studies. There were 300 replications for each experimental condition.

Data Simulation

The item discrimination (*a*) parameters for the primary dimension were sampled from a lognormal distribution with a mean of 1 and standard deviation of 0.5. To ensure the values were high enough but realistic, they were truncated between .4 and 2. After that, an angular distance was sampled from a uniform distribution with maximum 15° for tests with approximately simple structure or 15° and 45° for test with complex structure. The discrimination parameter of the secondary dimension was derived from the above two values. For each condition, a_1 was designated as primary dimension for the first half of the items with a_2 as secondary (e.g., item 1-10 for 20-item test) and for the second half a_2 was designated as primary with a_1 as secondary (e.g., items 10-20 for 20-item test).

All item difficulty parameters were sampled from a unit normal distribution. For the compensatory model there was one difficulty index while for the non-compensatory model there were two (one for each dimension).

All ability parameters were sampled from a bivariate normal distribution with mean of 0's and standard deviation of 1's. The correlation matrix for the ability space is defined as $\Sigma = \begin{Bmatrix} 1 & r \\ r & 1 \end{Bmatrix}$ where r is the correlation between two dimensions.

To address the indeterminacy in MIRT estimation, the discrimination parameter of two items (one representing each dimension) were fixed to zero for the secondary dimension. For example, fixing a_2 at 0 for the first item and a_1 at 0 for the 11th item in a 20-item test. In addition, indeterminacy was also addressed by fixing the scale with a mean = 0 and variance = 1 for the ability parameters.

To generate item responses, the item and ability parameters were placed into the compensatory model for Type I error studies or the non-compensatory model for power studies. The probability of a correct response was calculated. That probability was then compared to a value randomly generated from a uniform distribution between 0 and 1. If the probability was larger, the response was set equal to 1 (e.g., correct). Otherwise, the response was set to 0 (e.g., incorrect).

Lastly, the likelihood of a response pattern under the true model is calculated as

$$L(u_1, u_2, \ldots u_n|\theta) = \prod_{j=1}^{n} P_j^{u_j} Q_j^{1-u_j} \quad (22)$$

where u_j is the response to item j, P is the probability of a correct response and Q is the probability of an incorrect response (1-P). Because the values of likelihoods are so small due to the product function it is traditionally transformed using logarithms to create log-likelihoods.

To summarize, the steps in data generation were:

1. Generate item discrimination and difficulty parameters
2. Generate ability parameters

3. Calculate probability for the compensatory or non-compensatory model using generated parameters
4. Compare the probability to randomly generated value between 0 and 1 from a uniform distribution
5. If probability > random value, set the response to 1 (correct) or 0 otherwise
6. Calculate log-likelihoods for the true model

Model Estimation

flexMIRT was used for all model estimations. flexMIRT was introduced by Fraser and McDonald in 1986 (Ackerman et al., 2003; Maydeu-Olivares, 2001). MMLE was used for item parameter estimates and EAP for ability estimates. A benefit of flexMIRT is that is allows for non-normal latent trait distributions (Hortensius & Wang, 2014), is computationally fast, versatile, and user friendly. In addition, the program has been utilized successfully in studies with 2- and 3- parameter MIRT models (Paek & Cai, 2014). For these reasons this program was chosen as the compensatory estimation program for the current study.

The program used the generated response matrix to estimate the best fitting item and ability values. Those estimates were then output and steps 3 through 6 were repeated: the estimates were inserted into the compensatory model, the probabilities were generated, those compared against randomly generated values between 0 and 1 from a uniform distribution, and finally the log-likelihoods were calculated.

Convergence was assessed using winBUGS. flexMIRT convergence output was also monitored. Convergence is when a properly defined Markov Chain resolves into a stationary distribution which represents the joint posterior distribution of the model parameters. Even though convergence cannot be directly assessed, stationarity and run-length can be and are good

indications of convergence. For assessment purposes, two Markov Chains were run for each experimental condition to the point of convergence. Convergence was assessed using graphical (e.g., trace plots and autocorrelation functions) and statistical (e.g., Geweke, Heidelberger-Welch) diagnostics (Jackman, 2009). Based on stationarity stabilization at a maximum of 25,000 iterations, to be conservative, and based on previous research (Bolt & Lall, 2003; Yao, 2013; Yao & Boughton, 2007) a total of 30,000 iterations were run for each of the 300 replications per condition with a burn-in of 10,000.

Computation of Model Fit Statistics

The log-likelihoods from the true and estimated models were used in the calculations of both the Vuong and Clarke statistics. Both statistics use log-likelihood ratios in their calculations. This is one of the benefits of the Vuong and Clarke statistics and one of the many reasons they are often run as a pair. Another advantage is that they are fairly easy to compute.

The basic steps for the Vuong statistic are

1. Input log-likelihoods from generated (true) and estimated models
2. Calculate differences between the log-likelihoods from comparison models
3. Calculate mean and variance
4. Calculate probability under the unit normal distribution

The basic steps for the Clarke statistic are

1. Input log-likelihoods from the generated (true) and estimated models
2. Calculate differences between log-likelihoods from comparison models
3. Count positive and negative results
4. Calculate probability under binomial distribution

After all statistics were computed the null and power analyses were performed. The rejection rate was the dependent variables (DV) in all studies and power rates were used to define under which conditions the Vuong and Clarke statistic are applicable. Even though not all possible conditions could be examined in one study, the factors and levels being investigated in the current study aimed to represent those commonly encountered by practitioners.

CHAPTER 4: RESULTS

There are four sections in this chapter. The first section presents the descriptive statistics and model structure comparison results for the null conditions. The second section presents the same results but for the power conditions. The third section focuses on the outcomes of the ROC curve analyses. The final section details power levels and concludes with a quick summary of findings. In all sections the Vuong statistic results were presents first, followed by the Clarke statistic results.

Null Conditions

Recall that for the null conditions the true and estimated models are the same, the compensatory 2PL MIRT model. These models were examined for both the approximately simple and complex structures. Type I error level was examined at 0.05 but can easily be extended to other levels if necessary. A major focus of these results is to examine whether the sampling distribution of the fit statistics actually follow the assumed distributions, namely unit normal for Vuong statistic and binomial for Clarke statistic.

Table 2 presents the descriptive statistics for the Vuong statistic which include the mean of the statistic calculated using the 300 replication generated for each condition and the corresponding standard deviation. This will be the case with all successive tables, unless otherwise indicated. The correlation between traits is indicated by 'ρ'. A number of patterns

emerge in the table. First and also most strikingly, cell values are excessively large in the absolute sense, given that the sampling distribution is assumed be standard normal. Specifically, all values are much smaller than the -1.96 critical value at the .05 alpha level. Apparently, the assumption of the Vuong statistics following the standard normal distribution does not hold for these conditions. Second, cell values increase with test length and sample size. Accordingly, the values are largest for the 40-item, 1,000-subject conditions. Third, when sample size and length are fixed, the mean values are similar regardless of correlation between traits.

In addition, to examine whether the statistics behaved differently across test structure, *t*-tests were conducted to compare the approximately simple and complex structures. Skewness, kurtosis, histograms, and QQ plots were assessed for each distribution. Even though not centered around 0, t-test analyses were possible because all values were distributed normally around the mean for each condition. As shown in Table 3, overall, there was no difference between the two structures for the Vuong statistic. Only 2 out of the 16 conditions (ρ = .8, 20-item, 1,000-subject and the ρ = .3, 40-item, 1,000-subject) show significant difference but both have small effect size (.20 and .23 respectively). As both conditions are associated with a larger sample size, given the small effect sizes, significant *t*-tests may have been due to sample size rather than an actual treatment effect.

Table 4 gives the corresponding descriptive statistics for the Clarke statistic. As the Clarke statistic uses the total count to identify the number of comparisons where the true model was chosen, the count should be about half the total number of subjects (e.g., 250 for 500 subjects) for the null conditions. Similar to the Vuong statistic results, the counts for the Clarke statistic were quite small, which may also indicate that the assumption of a binomial distribution for this test is also tenable. However, clear patterns also emerge in the table. For instance, larger

sample sizes are associated with larger count values. Also, shorter tests are associated with larger counts regardless of the correlation levels between the two traits. Consequently, the 20-item, 1,000-subject conditions generated the largest counts for both structures. The *t*-test results for test structures in Table 5 give almost identical results as the Vuong statistic indicating that test structure has little impact on the behavior of the Clarke statistic, as well.

Table 2

Null Conditions, Descriptive Statistics for Vuong Statistic

Test Length (Items)	Sample Size (N)	Correlation Between Thetas ρ = 0	ρ = .3	ρ = .5	ρ = .8
		Approximately Simple Structure			
20	500	-16.64 (1.26)	-16.48 (1.19)	-16.51 (1.31)	-16.61 (1.40)
	1,000	-22.97 (1.50)	-22.68 (1.38)	-22.25 (1.40)	-21.63 (1.80)
40	500	-18.83 (0.99)	-18.62 (1.05)	-18.14 (1.02)	-17.34 (1.36)
	1,000	-26.12 (1.03)	-25.83 (1.11)	-25.09 (1.14)	-23.71 (1.31)
		Complex Structure			
20	500	-16.59 (1.24)	-16.52 (1.33)	-16.50 (1.28)	-16.81 (1.38)
	1,000	-22.95 (1.45)	-22.58 (1.28)	-22.35 (1.50)	-21.97 (1.58)
40	500	-18.81 (0.93)	-18.64 (0.98)	-18.17 (1.00)	-17.18 (1.45)
	1,000	-26.07 (1.13)	-25.58 (1.02)	-24.98 (1.05)	-23.73 (1.42)

Note. Standard deviations appear in parentheses bellow statistic means.

Table 3

Null Conditions, T-Test Results Comparing Test Structure for Vuong Statistic

Test Length (Items)	Sample Size (N)	Correlation Between Thetas			
		ρ = 0	ρ = .3	ρ = .5	ρ = .8
20	500	-0.43	0.33	-0.12	1.76
	1,000	-0.23	-0.95	0.75	2.40* (.20)
40	500	-0.26	0.18	0.43	-1.40
	1,000	-0.59	-2.99* (.23)	-1.16	0.15

Note. *=<.05; effect size Cohen's *d*; effect sizes appear in parentheses bellow significant *t*-test scores; all non-significant *t*-tests produced effect sizes <.10.

Table 4

Null Conditions, Descriptive Statistics for Clarke Statistic

Test Length (Items)	Sample Size (N)	Correlation Between Thetas			
		ρ = 0	ρ = .3	ρ = .5	ρ = .8
		Approximately Simple Structure			
20	500	99.27 (12.05)	100.83 (11.76)	100.24 (11.87)	99.09 (12.37)
	1,000	201.28 (22.44)	204.65 (19.55)	211.21 (19.62)	215.90 (23.11)
40	500	82.64 (8.39)	84.83 (8.99)	89.08 (9.19)	95.04 (10.41)
	1,000	160.64 (13.37)	165.98 (14.21)	175.64 (15.12)	193.26 (16.26)
		Complex Structure			
20	500	99.61 (11.46)	100.13 (11.83)	100.97 (14.34)	98.24 (11.62)
	1,000	200.78 20.31)	207.02 (19.33)	208.81 (11.97)	211.19 (20.58)
40	500	83.00 (8.30)	84.37 (8.22)	88.65 (20.58)	95.76 (10.80)
	1,000	162.07 (13.67)	170.14 (14.34)	176.83 (9.02)	192.35 (16.29)

Note. Standard deviations appear in parentheses bellow statistic means.

Table 5

Null Conditions, T-Test Results Comparing Test Structure for Clarke Statistic

Test Length (Items)	Sample Size (N)	Correlation Between Thetas			
		ρ = 0	ρ = .3	ρ = .5	ρ = .8
20	500	-0.35	0.72	-0.75	0.87
	1,000	0.28	-1.49	1.38	2.64* (.22)
40	500	-0.53	0.64	0.57	-0.83
	1,000	-1.29	-3.57* (.29)	-0.99	0.68

Note. * = <.05; effect size Cohen's *d*; effect sizes appear in parentheses bellow significant *t*-test scores; all non-significant *t*-tests produced effect sizes <.10.

Power Conditions

In contrast to the null conditions, the power study examined conditions that the true and estimated models are different. Specifically, the true model is the non-compensatory 2PL MIRT model and the estimated is its compensatory counterpart. The study focused on the power of the Vuong and Clarke statistics to detect the misfit between them. Like the null conditions, descriptive statistics were presented first along with test structure comparison. Since the underlying sampling distribution was not observed in the Type I error study, the receiver operating characteristic (ROC) curve results were provided to evaluate the overall effectiveness of the two statistics. Then the power rates based on Monte Carlo resampling were given.

Table 6 gives the descriptive statistics for the Vuong statistic. As expected now, all Vuong statistics are excessively large. Moreover, condition values increased with the correlation level between traits. They also increased with test length and sample size. For instance, the 20-item, 500-subject conditions have the smallest absolute values whereas the 40-item, 1,000-subject conditions have the largest absolute values. Other interesting patterns include that as

correlation between traits increase, the variances of the fit statistic increases. As test length and sample size increase, the variances decrease.

Results from comparing model structures are quite different from those under the null conditions. Table 7 shows that all but three conditions show significant differences between the two. While Cohen's *d* remained small for most conditions (< .10), medium effect (values around .5) is also observed. For instance, the 40-item, 1,000-subject condition show .4 and .5 level effect size. These results suggest that an awareness of the model structure is important in applying the statistic to detect misfit.

Presented in Table 8 are the descriptive statistics for the Clarke statistic. Again, mirroring the patterns from the null condition, as sample sizes increases, condition values increase. Another trend is shorter tests produce larger condition values. The combination of these two factors translates into the largest condition counts produced by the 20-item, 1,000-subject conditions and the smallest condition counts produced by the 40-item, 500-subject conditions. Interestingly, the variance is still consistently larger with larger sample size, but it fluctuates across correlation and sample size.

In comparing model structure, all but five conditions in Table 9 show significant differences between the two test structures. The patterns parallel those of the Vuong statistic in that Cohen's *d* remains small for most conditions (< .10) with medium effect (values around .5) for some conditions, Again, the 40-item, 1,000-subject condition has the largest effects. These results indicate that the two statistics respond in similar fashion to a change in model structure and an awareness of this is important in applying the statistic.

Table 6

Power Conditions, Descriptive Statistics for the Vuong Statistic

Test Length (Items)	Sample Size (N)	Correlation Between Thetas			
		$\rho = 0$	$\rho = .3$	$\rho = .5$	$\rho = .8$
		Approximately Simple Structure			
20	500	-12.43	-13.39	-14.50	-16.93
		(1.95)	(2.23)	(2.29)	(2.82)
	1,000	-14.58	-16.13	-17.26	-20.56
		(1.72)	(2.89)	(3.19)	(3.31)
40	500	-12.70	-13.27	-13.81	-16.32
		(1.15)	(1.16)	(1.28)	(1.33)
	1,000	-16.36	-16.70	-17.23	-20.33
		(1.18)	(1.24)	(1.23)	(1.41)
		Complex Structure			
20	500	-12.43	-13.70	-15.08	-16.73
		(1.59)	(2.25)	(2.38)	(2.34)
	1,000	-14.95	-16.36	-18.09	-21.40
		(1.77)	(2.48)	(3.17)	(4.02)
40	500	-13.01	-13.46	-14.13	-16.74
		(1.09)	(1.20)	(1.12)	(1.31)
	1,000	-16.74	-17.02	-17.81	-21.08
		(1.27)	(1.23)	(1.34)	(1.48)

Note. Standard deviations appear in parentheses bellow statistic means.

Table 7

Power Conditions, Comparing Test Structure for Vuong Statistic

Test Length (Items)	Sample Size (N)	Correlation Between Thetas			
		ρ = 0	ρ = .3	ρ = .5	ρ = .8
20	500	0.01	1.72	3.01* (.25)	2.77* (.21)
	1,000	2.60* (.21)	1.03	3.19* (.26)	3.91* (.23)
40	500	3.38* (.20)	2.09* (.16)	3.34* (.27)	6.41* (.32)
	1,000	3.84* (.31)	3.24* (.26)	5.51* (.44)	2.77* (.52)

Note. * = <.05; effect size Cohen's *d*; effect sizes appear in parentheses bellow significant *t*-test scores; all non-significant *t*-tests produced effect sizes <.10.

Table 8

Power Conditions, Descriptive Statistics for Clarke Statistic

Test Length (Items)	Sample Size (N)	Correlation Between Thetas			
		ρ = 0	ρ = .3	ρ = .5	ρ = .8
		Approximately Simple Structure			
20	500	139.50 (18.21)	130.62 (19.14)	122.13 (18.72)	100.15 (20.33)
	1,000	317.93 (24.84)	297.25 (38.44)	281.44 (40.54)	236.61 (39.28)
40	500	134.85 (11.81)	129.78 (10.90)	125.11 (11.86)	105.87 (10.99)
	1,000	290.59 (17.25)	284.64 (17.32)	276.92 (17.39)	238.47 (18.26)
		Complex Structure			
20	500	138.54 (15.74)	128.70 (19.12)	116.70 (19.41)	101.58 (18.07)
	1,000	313.90 (26.69)	292.85 (34.41)	271.43 (39.75)	227.02 (45.37)
40	500	132.14 (10.93)	128.65 (10.79)	123.14 (11.15)	100.65 (10.71)
	1,000	285.35 (18.21)	280.17 (18.17)	269.61 (19.12)	229.42 (18.22)

Note. Standard deviations appear in parentheses bellow statistic means.

Table 9

Power Conditions, Comparing Test Structure for Clarke Statistic

Test Length (Items)	Sample Size (N)	Correlation Between Thetas ρ = 0	ρ = .3	ρ = .5	ρ = .8
		Complex Structure			
20	500	0.69	1.23	3.48* (.29)	-0.91
	1,000	2.30* (.16)	1.48	3.05* (.25)	2.77* (.23)
40	500	2.91* (.23)	1.28	2.10* (.17)	5.89* (.48)
	1,000	3.61* (.30)	3.09* (.25)	4.90* (.40)	6.08* (.50)

Note. * = <.05; effect size Cohen's *d*; effect sizes appear in parentheses bellow significant *t*-test scores; all non-significant *t*-tests produced effect sizes <.10.

ROC Results

In the current study the ROC curve was used to assess the ability of the Vuong and Clarke statistics to detect misfit between non-compensatory and compensatory models. Recall that an AUC value above .9 is considered exceptional discrimination, or the measure is highly effective.

Tables 10 and 11 give the AUC values for the Vuong and Clarke statistics, respectively. In general, longer tests and larger sample sizes result in larger AUC values while higher correlations are associated with smaller AUC values, regardless of test structure. Based on the patterns revealed in the descriptive statistics, the AUC patterns are somewhat expected and not surprising. Looking at specific AUC values for the Vuong statistic, out of 32 total values, 9 had values of 1.0 (perfect discrimination) and 15 had values above .9 (exceptional discrimination). Only 4 of the AUC values were less than the acceptable .8 cutoff. For both structures, the

smallest values were produced by the 20-item, 500-subject condition with correlations of $\rho = .5$ and $\rho = .8$. For the Clarke statistic, of the 32 AUC values, 9 had values of 1.0 (perfect discrimination), 15 had values above .9 (exceptional discrimination), and 6 were less than the acceptable .8 cutoff.

Overall both the Vuong and Clarke statistics produced excellent AUC values, providing strong evidence that both statistics have the potential prowess to discriminate between compensatory and non-compensatory models under the examined conditions. However, there are some exceptions. First, regardless of test structure, when the correlation between traits reach .8 the AUC values generally decrease to an unacceptable level. Even at the .5 correlation level, AUC values are still low with short test (e.g., 20-items) and small sample size (e.g., 500-subjects). In comparing the Vuong and Clarke statistics, the AUC values are actually quite similar. When the values are different, the Vuong appears to have slightly larger values for the shorter test with approximately simple structure while the Clarke appears to have slightly larger values for the longer test. For tests with complex structure, the Clarke statistic generally shows larger AUC values.

Table 10

AUC Values for Vuong Statistic

Test Length (Items)	Sample Size (N)	Correlation Between Thetas			
		ρ = 0	ρ = .3	ρ = .5	ρ = .8
		Approximately Simple			
20	500	.95	.90	.70	.72
	1,000	.99	.96	.92	.73
40	500	1.0	1.0	.99	.74
	1,000	1.0	1.0	1.0	.96
		Complex			
20	500	.98	.88	.76	*.56*
	1,000	.99	.97	.90	*.69*
40	500	1.0	.99	.99	*.62*
	1,000	1.0	1.0	1.0	.92

Note. Italicized results are below acceptable discrimination ability cutoff values.

Table 11

AUC Values for Clarke Statistic

Test Length (Items)	Sample Size (N)	Correlation Between Thetas			
		ρ = 0	ρ = .3	ρ = .5	ρ = .8
		Approximately Simple			
20	500	.95	.91	.84	*.57*
	1,000	.99	.96	.93	.73
40	500	1.0	.99	.99	.79
	1,000	1.0	1.0	1.0	.96
		Complex			
20	500	.98	.90	.78	*.59*
	1,000	.99	.97	.91	.71
40	500	1.0	1.0	.99	*.64*
	1,000	1.0	1.0	1.0	.94

Note. Italicized results are below acceptable discrimination ability cutoff values.

Power Results

The ROC analysis results provide extremely strong evidence that the Vuong and Clarke statistics should be able to discriminate the compensatory and non-compensatory MIRT models. In response, a Monte Carlo resampling technique was performed to calculate the power of the two statistics. As described previously in the Methods section, the cut point that equals the 97.5% upper limit of the null distribution was used to calculate the power in the alternative distribution. The cut off points are presented in Tables 12 and 13. For example, for the condition with .3 correlation, 20-items, 500-subjects, and approximately simple structure, the null distribution of the Clarke statistic has a 97.5% upper limit cut point of 124, comparing that value to the alternative distribution results a rejection rate of .67, a fairly low power as it implies that the Clarke statistic can identify model data misfit only 67% of the time.

The power levels for the Vuong statistic are presented in Table 14. In general, the power is quite high. Most of the conditions have acceptable to excellent (.80 - .99) or perfect power (1.0), yet there is a grouping with low power. In about one third of conditions, the statistic enjoys perfect or near perfect power. On the other hand, it is also clearly apparent that power is lacking for the .8 correlation conditions.

Interestingly, the pattern of low and high power is the same for both test structures. As can be seen in the table, the differentiation between the high and low power groupings is very clear. Specifically, the conditions with high power have no correlation or low correlation between traits. Also, higher power is associated with larger sample size and longer test length. On the other hand, there is low power when correlation is .8, regardless of test length or sample size. Power also lacks for 20-item, 500-subject conditions where correlation is .3 or .5.

The Clarke statistic power levels are presented in Table 15. Like the Vuong statistic, the Clarke statistic has excellent power. Under 10 conditions power is perfect. On the other hand, there are also 8 low power conditions (.04 - .67). In general, the patterns are similar to what has been previously observed for Vuong: increased power with increased sample size, increased power with increased test length, and decreased power with increased correlation between traits, especially when correlation is .8. Also in line with previous results is a clear grouping of low and high power conditions. The departure from these patterns is that the power levels are different depending on test structure. For the Clarke statistic, the power actually increases with complex structure. This is most noticeable in that 2 additional complex structure conditions have perfect power and 2 more have high power when the same conditions for approximately simple structure have low power ($\rho = .3$, 20-item, 500-subject and $\rho = .5$, 20-item, 500-subject).

A direct comparison of the power for the Vuong and Clarke statistics is summarized in Table 16. Table conditions identified with 'Vuong' are conditions where the Vuong statistic outperformed the Clarke statistic. Likewise, conditions identified with 'Clarke' are where the Clarke statistic outperformed. In addition, the conditions with a 'Low' designation mean neither test produced a satisfactory power level, conditions with 'Equal' mean both tests produced acceptable power (.8-.99), and finally, conditions with a 'Perfect' designation mean both statistics produced 1.0 power level.

Some interesting facts appear in the table. Most apparent is that under most conditions the Vuong statistic, the Clarke statistic, or both statistic are able to detect misfit. In fact, the statistics produced that same power level for 40% of the conditions. Even though there are similarities across the structures, for practical purposes, which statistic or statistics are best at detecting misfit depends on test structure.

For tests with approximately simple structure, there are 11 conditions where misfit is being detected and only 5 where it is not. Specifically, there are 3 conditions where the tests have adequate to excellent power and 4 conditions where there is perfect power. This indicates that both the Vuong and Clarke statistics detect misfit for these conditions. However, the Vuong statistic outperforms with greater power for 2 of the 20-item conditions. This indicates that when not performing equally, the Vuong statistic detects misfit better than the Clarke with some shorter item conditions. The Clarke test outperforms only once and that is under a 40-item condition. There are 5 conditions where the tests produce equally low power. As discussed previously, low power is observed especially when correlation is high.

For tests with complex structure, misfit is identified for the majority of conditions (12 out of 16). Specifically, there are 2 conditions where the Vuong and Clarke statistics have equal

adequate power and 5 conditions with perfect power. In addition, and quite interestingly, the Clarke statistic outperformed the Vuong statistic on multiple occasions whereas only once did the Vuong outperform by a slight margin. These results indicate that the Clarke statistic may be better suited for identifying misfit for tests with complex structure.

In summary, the investigation into the Vuong and Clarke statistics has generated consistent and informative findings that attest to their ability to detect misfit between the compensatory and non-compensatory models. Even though the assumed distributions were not observed, the ROC analyses found that the statistics have excellent discriminating abilities and high power can be attainted, which in turn was confirmed by deriving the power by Monte Carlo resampling. Other than a small group of specific conditions, power is quite high. Taking into consideration the patterns of results, it appears that all factors examined including test structure, sample size, test length, and correlation between traits were important and had shown an impact on the performance of the fit statistics.

Table 12

Power Cutoff Values for Approximately Simple Structure for Vuong and Clarke Statistics

Test Length (Items)	Sample Size (N)	Correlation Between Thetas			
		ρ = 0	ρ = .3	ρ = .5	ρ = .8
		Vuong statistic			
20	500	-14.00	-14.04	-13.60	-13.66
	1,000	-19.70	-19.70	-19.63	-17.33
40	500	-16.81	-16.41	-16.01	-14.53
	1,000	-23.96	-23.59	-23.01	-20.96
		Clarke statistic			
20	500	125	127	128	124
	1,000	243	248	249	252
40	500	100	101	105	117
	1,000	136	201	205	233

Table 13

Power Cutoff Values with Complex Structure for Vuong and Clarke Statistics

Test Length (Items)	Sample Size (N)	Correlation Between Thetas			
		ρ = 0	ρ = .3	ρ = .5	ρ = .8
		Vuong statistic			
20	500	-14.03	-13.85	-13.71	-14.02
	1,000	-19.84	-19.80	-19.60	-18.46
40	500	-17.02	-16.91	-16.38	-14.01
	1,000	-23.51	-22.98	-22.87	-20.21
		Clarke statistic			
20	500	124	125	123	124
	1,000	249	245	256	260
40	500	101	103	107	120
	1,000	189	193	205	224

Table 14

Power for Vuong Statistic

Test Length (Items)	Sample Size (N)	Correlation Between Thetas			
		ρ = 0	ρ = .3	ρ = .5	ρ = .8
		Approximately Simple			
20	500	.86	.71	.41	.05
	1,000	.99	.93	.88	.05
40	500	1.00	.90	.96	.06
	1,000	1.00	1.00	1.00	.68
		Complex			
20	500	.84	.65	.30	.06
	1,000	.99	.98	.91	.17
40	500	1.00	1.00	.96	.01
	1,000	1.00	1.00	1.00	.23

Table 15

Power for Clarke Statistic

Test Length (Items)	Sample Size (N)	Correlation Between Thetas			
		ρ = 0	ρ = .3	ρ = .5	ρ = .8
		Approximately Simple			
20	500	.86	.67	.55	.07
	1,000	.99	.93	.83	.27
40	500	1.00	.97	.95	.07
	1,000	1.00	1.00	1.00	.80
		Complex			
20	500	.81	.92	.82	.31
	1,000	.99	.99	.92	.04
40	500	1.00	1.00	1.00	.44
	1,000	1.00	1.00	1.00	.92

Table 16

Power

Test Length (Items)	Sample Size (N)	Correlation Between Thetas			
		ρ = 0	ρ = .3	ρ = .5	ρ = .8
		ASST			
20	500	E	Vuong	L	L
	1,000	E	E	Vuong	L
40	500	P	Clarke	E	L
	1,000	P	P	P	L
		Complex			
20	500	Vuong	Clarke	L	L
	1,000	E	Clarke	E	L
40	500	P	P	Clarke	L
	1,000	P	P	P	Clarke

Note. E = equal power for both statistics but not perfect; P = perfect power for both statistics; L = low power for both statistics less than .6; Vuong = Vuong statistic has higher power than Clarke statistic; Clarke = Clarke statistic has higher power than Vuong statistic.

CHAPTER 5: DISCUSSION

This study was set up to investigate whether the Vuong and Clarke statistics can be used to detect model-data fit for compensatory MIRT models. Its importance is three fold. First, currently no statistics can provide probabilistic statements about model-data fit between the non-nested compensatory and non-compensatory MIRT models. Second, the Vuong and Clarke statistics have been utilized prolifically for non-nested in economic and political science and have great potential to contribute to educational measurement. Third, good model fit statistics will not only reduce the damage of using wrong MIRT models but also promote the use of less known models, in this case, the non-compensatory models.

Results from the simulation study provide strong support for using both the Vuong and Clarke statistics for detecting model-data fit for MIRT models. There are numerous conditions where both statistics can detect misfit with ease. When sample size reaches 1,000, regardless of test structure and test length, power is very high for low and medium correlated traits. This sample size requirement can easily be fulfilled for educational tests. For instance, state tests usually have over 1,000 test takers for each grade in large school districts. Meanwhile, as correlation between cognitive traits (e.g., reading, mathematics, and science) is generally medium level, applying these two statistics will help researchers study whether students are used the target traits in a compensatory manner or not.

Another condition where both statistics work really well is when the person traits are uncorrelated. Power is very high for both statistics regardless of test structure, sample size, or test length. In practice, low correlations happen more often between psychological traits than between educational traits. In situations like these the fact that power is high even at smaller sample sizes is very helpful, given studies in psychology generally have smaller sample sizes.

Results also clearly indicate when these two statistics are incapable of detecting the misfit. Once the correlation between two traits reach .8, the Vuong statistic will not be able to detect the difference between the compensatory and non-compensatory relationships. The Clarke test can do better but only when both sample size and test length are large. Also, for correlations of .3 and .5, the statistics do not detect differences with 20-items, 500-subjects. Either one or both of the number of items or subjects need to increase for detection to occur.

This research examined the impact of four test characteristics on the performance of the two fit statistics. The results show clear patterns for all of them, of which practitioners should be aware.

For test structures, there was a distinct difference in the performance of the statistics between the structures, particularly for the Clarke statistic. The Clarke statistic produces higher power rates for tests with complex structure than those with approximately simple structure. One possible reason for the increase in performance is if the underlying log-likelihood distribution for these complex structure conditions is leptokurtic. When this is the case, the Clarke statistic performs particularly well. Another possible explanation is that items in tests with complex structure have a larger discrimination on the secondary dimension, which tends to make the two models more distinct. In practice, as the true test structure is usually unknown, using the Clarke statistic will detect misfit more often than the Vuong statistic when tests have complex structure.

On correlation between traits, power decreases with the increase of the correlation. The largest power rates are when there is no correlation. The smallest power rate for no correlation is .86, which is still fairly large, so the statistics maintain power with even smaller sample sizes and test lengths. It would be interesting to investigate how small the sample size and test length can be for the high power to hold. The lowest power is observed when correlation is .8. One possible

reason for the lack of power is that those tests are more unidimensional than multidimensional, hence most fit statistics will be not able to tell the difference between multidimensional models. This is consistent with previous studies that found that with increased correlation, compensatory, and non-compensatory models became indistinguishable (Spray et al., 1990). Clarke (2003) also found that power decreases with increased correlation.

With increasing test length, as expected, the ability to detect misfit by both statistics increases. Longer test lengths increase the accuracy and consistency of the ability and item parameter estimates, which in turn improves the effectiveness of model fit statistics. The test lengths of 20- and 40-items represent short to medium and long lengths and are common in research and practice. Since the statistics are able to detect misfit for many of the 40-item conditions, they should also be able to detect misfit for longer tests. However, practitioners should be more cautious with 20-item tests. Contingent on the levels of other test characteristics, one or both of the statistics will not be able to detect the misfit.

As with test length, power also increases with increased sample size, which is not surprising for the likelihood-based tests. As long as correlation is not high, both statistics will detect power when the sample size is at or above 1,000. As there are situations where sample size can be quite small, such as data analysis at the school level, it is necessary to investigate the power of the statistics for conditions with sample sizes less than 500. Clarke (2010) found that the statistics were able to detect misfit with sample sizes as small as N=50.

As observed in previous research (Clarke, 2007), the two statistics perform noticeably different. While they do perform equally in many conditions, there are also conditions where one outperforms the other. For instance, for tests with approximately simple structure the Vuong statistic performs better for 20-item conditions. The Clarke statistic performs better for tests with

complex structure. The reason may lie in the nature of the statistics themselves. To review, the Clarke statistic was proposed to address some of the inherent shortcomings of the Vuong statistic, such as lack of power when the distribution of the log-likelihood ratios is not normal. For complex structure, if the underlying log-likelihood distributions are not normal it could explain why the Clarke statistic is better at detecting misfit. For the Vuong statistics, with approximately simple structure and shorter tests, the distributions may be more standard normal. Without definite knowledge of the test structure, the previous recommendation to run them simultaneously also applies here, especially when test length is short.

Yet even with the pattern of results supporting the potential of both statistics, there are points of interest to be discussed and resolved prior to wide-spread application. First is the large relative values for the Vuong and Clarke statistics. The statistics were expected to produce values consistent with a standard normal or binomial distributions (e.g., Vuong and Clarke, respectively) which would result in appropriate reject rates depending on study. For instance, less than 5% for the Type I error studies. Instead, the rejection rate was 100% due to the large relative values of the statistics. One possible reason for this discrepancy is these statistics were not designed for latent trait models, like IRT models. The Vuong and Clarke statistics have been primarily used to compare models using observable manifest traits. The difference between manifest and latent trait models could potentially cause misalignment and issues in parameter estimation which include lack of convergence in equation solutions and increased measurement error (Fox, 2008; Hambleton & Swaminathan, 2010; Wang, 2015). The large number of latent trait parameters may also contribute to these issues (Hambleton & Cook, 1977; Hambleton et al, 1978). It is worth mentioning that two adjustments have been suggested for the Vuong statistic,

one addressing latent trait issues for nested models and the other addressing possible issues with traditional non-nested models.

The first adjusted Vuong test is the Vuong-Lo-Mendell-Ruben statistic (VLMR; Lo et al, 2001) for use with nested and overlapping latent-trait models. With the adjustment, studies have found that convergence and accuracy in identifying misfit with nested models was improved versus the traditional Vuong statistic and information based fit indices including AIC and BIC (Lo et al, 2001; Morgan et al, 2016). The second adjusted Vuong test is the nondegenerate Vuong statistic (Shi, 2015) from a recent study investigating the traditional Vuong statistic as applied to traditional non-nested models. The results indicate that the Vuong statistic over-rejected when the log-likelihood ratios exceeded normal critical levels with the distribution being between a normal and nonstandard distribution (Shi, 2015). The author proposed an adjusted Vuong test, coined the nondegenerate Vuong test, which through a series of corrections addresses possible bias in the numerator and possible randomness in the denominator to curtail the over-rejection rates. The availability of these adjusted statistics suggest that the Vuong specifically may need some type of correction for use with IRT models. Currently, there are no Clarke statistic adjustments in the literature pertaining to latent-trait models. One possible reason is that the Clarke statistic was specifically created for non-nested comparisons, which is less common than the nested comparison. But for situations like the current study, it is worth looking into that option, as well.

Second, because the Type I error under the assumed sampling distribution was not observed, a Monte Carlo resampling technique was employed in this study. In practice, it would be convenient to give practitioners cut points, like the 1.96 from the standard normal distribution. One issue for such a practice for these two fit statistics is that the cut point will be different for

each condition, which means every possible combination of factors would need to be examined and a cut point provided for each. A more feasible while less convenient approach is to use the Monte Carlo resampling method employed in this study.

As a practical example, a practitioner has data for a test with 40-multidimensional items taken by 1,500-subjects. The test is assessing geometric and algebraic knowledge in a school district. Rather than assuming a compensatory mathematical and theoretical fit to the data, the practitioner wants to test whether the data fits a compensatory or non-compensatory model best. Previously, because these models are non-nested, the only options available were fit indexes like AIC and BIC from which no probabilistic statements can be made. Since the current study revealed that both the Vuong and Clarke statistics have almost perfect power in detecting the misfit, practitioners may choose to apply them to the data. Since the sample size is larger than 1,000, the observed highest power rates from this study should be applicable. The data would be used with the compensatory and non-compensatory models to generate fit statistics. Those statistics would be evaluated by comparing them to a derived empirical sampling distribution. The empirical sampling distribution is created by using the compensatory model to obtain item parameter estimates, running Monte Carlo resampling based on those estimates, and then accumulate an empirical sampling distribution of the fit statistic comparing the two compensatory model. This is not a convenient process and limits the application of the statistics as the knowledge of Monte Carlo techniques is required.

As mentioned, the Vuong and Clarke statistics are proposed to be better model-data fit options to the traditional fit indexes of AIC and BIC. The current study did not present AIC and BIC values because the indexes would experience the same offset values observed with the Vuong and Clarke statistics. The reason is the large differences between the log-likelihoods

between the compared models. These same log-likelihoods are used to calculate the fit indexes. In addition, the focus of the study was on the performance of the Vuong and Clarke statistics and not as a comparison to AIC and BIC, although this could be done in future research.

Limitations and Further Research

This study suffers from a number of major limitations. First, a Monte Carlo simulation like this does not provide an analytical solution to the general question of whether the fit statistics can be used for the intended model. In other words, its findings can only be applied to the conditions investigated in the study, which by nature, is limited. Many important conditions may have been left out. In this study, the investigated conditions were restricted to two dimensions, two structures, two sample sizes, two test lengths, and four correlation levels between traits. The factors and levels were chosen to simulate the most common testing conditions practitioners may encounter. It is recommended that future research expand to include more levels of each factor, such as higher dimensions, longer tests, and larger sample sizes.

The second limitation is the convenience of suggested methods. Even with faster computers today, the time necessary to compute the fit statistics is quite long, due to the resampling process. Even with 20-items and 500-subjects it takes about three minutes to compute one replication of the fit statistics. In the era of click and run, this is quite a long time for practitioners. This time requirement also restrained the number of conditions that could be included in the current study.

Finally, another important limitation, not by design but revealed during the simulation study, is the accuracy of ability estimation for the MIRT models. In the current study, correlations between the true and estimated thetas ranged from .70 to .80. These relatively low correlations may have led to less optimal performance of the fit statistics and to large differences

in the log-likelihoods. To investigate this problem, three computer programs were run and checked. Two programs, BMIRT (Yao, 2003) and winBUGS (Lunn et al, 2000), produced correlations around .70. One program, flexMIRT (Cai, 2013), produced the largest correlation of .80. Based on this fact and others (consistency, speed, long history of application, and validation; Han & Paek, 2014) flexMIRT was used for this study. Future research should explore new ways to improve ability estimation accuracy.

There are many directions for further research. First, expand the testing conditions to include more levels of ability dimension, test length, sample size, test structure, and correlation to create a comprehensive guideline for use of the Vuong and Clarke statistics.

Second, investigate whether good rules of thumb can actually be derived. This includes providing more evidence that the Vuong statistic be the preferred for all shorter tests and the Clarke statistic be the preferred for longer tests with approximately simple structure. Also, for complex structure, that the Clarke statistic be the preferred and default choice over the Vuong statistic.

Third, examine possible mathematical adjustments to the statistics as suggested by the discussion of the VLMR and nondegenerate Vuong statistics. Each adjustment addresses different issues associated with non-nested and latent trait models that potentially impact the performance of the Vuong statistic. An exploration into the adjustments and if they modify the statistic in a manner that improves performance in the current study is suggested.

Fourth, compare the performance of the Vuong and Clarke statistics to the traditional non-nested model indexes of AIC and BIC. If the Vuong and Clarke statistics identify misfit more accurately, it would provide proof that they are an improvement and should be applied versus the other options.

Finally, evaluate estimation programs. This includes comparing estimation programs for ability parameter accuracy and recovery for compensatory models as well as non-compensatory models. The programs that provide non-compensatory estimations are relatively new. Once their effectiveness is established, one can compare the true compensatory to the estimated non-compensatory, which would further expand the knowledge of the discrimination ability of the statistics by providing a complete picture of their usability.

Conclusions

In summary, this study provided strong evidence of the Vuong and Clarke statistics as viable options for detecting model-data fit between compensatory and non-compensatory MIRT models. A series of studies revealed high power rates for both statistics for many testing conditions. The statistics were especially powerful when sample size and test length were large and correlation between traits small. The Vuong and Clarke statistics performed differentially for some conditions, as expected. For example, the Clarke statistics performed better with complex structure tests. Also, because there were clear patterns of results, extrapolation beyond the current condition matrix is possible. When applying the statistics, care and awareness of the shortcomings of the current study should be taken into consideration. The great potential of the Vuong and Clarke statistics in the area of educational measurement was also recognized. Overall, these findings show that the Vuong and Clarke statistics can be used to detect misfit for compensatory MIRT models.

REFERENCES

Ackerman, T. A. (1992). A didactic explaination of item bias, item impact, and item validity form a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*(4), 255-278.

Ackerman, T., Gierl, M., & Walker, C. (2003). An NCME instructional module on using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37-53.

Akaike, H. (1974). A new look at the statistical model identification. *Free Transactions on Cutomatic Control, 19*(6), 716-723.

Allen, D. D., & Wilson, M. (2006). Introducing multidimensional item response modeling in health behavior and health education research. *Health Education Research, 21*(1), i73-i84.

Andersen E. B. (1982). Latent trait models and ability parameter estimation. *Applied Psychological Measurement, 6*(4), 445-461.

Babcock, B. (2011). Estimating a noncompensatory IRT model using metropolis within Gibbs sampling. *Applied Psychological Measurement, 35*(4), 317-329.

Beguin, A., & Glas, C. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66*(4), 541-562.

Bolt, D. M. (2001). Conditional covaariance-based representation of multidimensional test struccture. *Applied Psychologial Measurement, 25*(3), 244-257.

Bolt, D., & Lall, V. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Morkov chain Monte Carlo. *Applied Psychological Measurement, 27*(6), 395-414.

Cai, L. (2013). flexMIRT version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Clarke, K. A. (1998). Nonnested model testing for world politics assessing binary choice models. *American Political Science Association Conference*. Boston, MA.

Clarke, K. A. (2001). Testing nonnested models of international relations: Reevaluating realism. *American Journal of Political Science, 45*(3), 724-744.

Clarke, K. A. (2003). Nonparametric model discrimination in international relations. *Journal of Conflict Resolution, 47*(1), 72-93.

Clarke, K. A. (2007). A simple distribution-free test for nonnested model selection. *Political Analysis, 15*(3), 347-363.

Clarke, K. A., & Signorino, C. S. (2010). Discriminating methods: Tests for non-nested discrete choice models. *Political Science, 38*(2), 368-388.

Croker, L., & Algina, J. (1986). *Introduction to classical and & modern test theory*. Orlando, FL: Harcourt, Brace, & Javanovich, Inc.

Czado, C., Schabenberger, H., & Erhardt, V. (2014). Non-nested model selection for spatial count regression models with application to health insurance. *Statistical Papers, 55*(2), 1-22.

de la Torre, J., & Patz, R. (2002). A multidimensional item response theory approach to simultaneous ability estimation. *Annual Meeting of the NCME* (pp. 1-16). New Orleans, LA: National Council on Measurement in Education.

DeAyala, R. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

DeMars, C. E. (2004). Type I errors rates for generalized graded unfolding model fit indices. *Applied Psychological Measurement, 28*(1), 48-71.

Elliot, A. C., & Woodward, W. A. (2010). *SAS Essentials: Mastering SAS for research*. San Fransisco, CA: Jossey-Bass Publications.

Finch, H. (2011). Multidimensional item response theory parameter estimation with nonsimple structure items. *Applied Psychological Measurement, 35*(1), 67-82.

Fox, J. (2008). *Applied regression analysis and generalized linear models*. Thousand Oaks, CA: SAGE Publications.

Genius, M., & Strazzera, E. (2001). *Model selection and tests for non-nested contingent valuation models: An assessment of methods*. Milan, Italy: Fondazione Eni Enrico Mattei.

Hall, A. R., & Pelletier, D. (2011). Non-nested testing in models estimated via generalized method of moments. *Econometric Theory, 27*(2), 443-456.

Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement, 14*(2), 75-96.

Hambleton, R. K., & Swaminathan, H. (2010). *Item response theory: Principles and applications*. Norwell, MA: Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Gifford, E., & Gifford, J. A. (1978). Developments in latent trait theory: Models, technical issues, and applications. *Review of Educational Resaearch, 48*(4), 467-510.

Hambleton, R. K., Swaminathan, H., & Roger, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE Publications.

Han, K. T., & Paek, I. (2014). A review of commercial software packages for multidimensional IRT modeling. *Applied Psychological Measurement, 38*(6), 486-498.

Hartig, J., & Hohler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation, 35*(2), 57-63.

Hong, H., & Preston, B. (2006). Nonnested model selection criteria. *Working paper, Duke University*.

Hortensius, L., & Wang, C. (2014). *Practical guidelines for the estimation of multidimensional ordered polytomous models*. Presentation at NCME.

Hulin, C. L., Parsons, C. K., & Drasgow, F. (1995). *Item response theory: Application to psychological measurement*. Burr Ridge, IL: Irwin Professional Publishing.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. West Sussex, England: John Wiley & Sons, Ltd.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*(3), 767-778.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: SAGE Publications.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Publishing Co.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS: A Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing, 10*, 325-337.

Mair, P., Reise, S. P, & Bentler, P. M. (2011). IRT Goodness-of-fit using approaches from logistic regression. Department of Statistics Papers, pub date 10/25/2011 Dept of Statistics, UCLA. http://escholarship.org/uc/item/76b7682n

Maydeu-Olivares, A. (2001). Multidimentional item response theory modeling of binary data: Large sample properties of NOHARM estimates. *Journal of Educational and Behavioral Statistics, 26*(1), 51-71.

McKinely, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*(1), 49-57.

Mooney, C. (1997). *Monte Carlo simulation.* Iowa City, IA: Sage Publications.

Morgan, G. B., Hodge, K. J., & Baggett, A. R. (2016). Latent profile analysis with nonnormal mixtures: A Monte Carlo examination of model selection usign fit indices. *Computational Statistics & Data Anslysis, 93*, 146-161.

Orlando, M., & Thissen, D. (2000). Likelihood based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50-64.

Osteen, P. (2010). An introduction to using multidimensional item response theory to assess laten factor structures. *Journal of the Society for Social Work and Research, 1*(2), 66-82.

Oaek, I., & Cai, L. (2014). A comparison of item parameter standard error estimation procedures for unidimensional and multidimensional item response theory modeling. *Educational and Psychological Measurement, 74*(1), 58-76.

Pesaran, M. H., & Deaton, A. S. (1978). Testing non-nested nonlinear regression models. *Ecomentrica: Journal of the Econometric Society, 46*(3), 677-694.

Pesaran, M. H., & Ulloa, R. D. (2008). Non-nested hypothesis.In Durlauf, S. & Blume, L. E. (Eds.), *The New Palgrave Dictionary of Economics* (2nd ed., Vol. 2, pp. 609-642). London, England: Palgrave MacMillan.

Ranger, J., & Kuhn, J. (2014). Testing fit of latent trait models for responses and response times in tests. *Psychological Test and Assessment Modeling, 56*(4), 382-404.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*(4), 401-412.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25-36.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY: Springer.

Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika, 66*(1), 79-97.

Shi, X. (2015). A nondegenerate Vuong test. *Quantitative Economics, 6*(1), 85-121.

Spray, J., Davey, T., Reckase, M., Ackerman, T., & Carlson, J. (1990). *Comparison of two logistic multidimensional item response theory models.* Iowa City, IA: ACT Inc.

Stone, C. A. (2000). Monte Carlo based null distribution for and alternative hoodness-of-fit test statistic in IRT models. *Journal of Educational Measurement, 37*(1), 58-75.

Stout, W., Habing, B., Douglas, J., Kim, H., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*(4), 331-354.

Svetina, D. (2012). Assessing dimensionality of noncompensatory item response thoery with complex structures. *Educational and Psycholgical Measurement, 73*(2), 312-338.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *proceedings of the 1977 computerized adaptive testing conference* (pp. 82–98). Minneapolis, MN: University of Minnesota.

Traub, R. (1997). Classical test theroy in historical perspective. *Educational Measurement: Issues and Practice, 16*(4), 8-14.

Vollema, M. G., & Hoijtink, H. (2000). The multidiemsionalty of self-report schizotypy. *Schizophrenia Bulletin, 26*(3), 565-575.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, *57*, 307-333.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement*, *38*(2), 147-163.

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as diagnostic aid. *Journal of Educational Measurement*, *40*(3), 255-275.

Wang, C. (2015). On latent trait estimation in multidimensional compensatory response models. *Psychometrika, 80*(2), 428-449.

Yao, L. (2003). BMIRT: Bayesian multivariate item response theory [computer software]. Monterey, CA: CTB/McGraw-Hill.

Yao, L. (2013). *Classification accuracy and consistency indices for summed scores enhanced using MIRT for test of mixed item types.* Monterey Bay, CA: Defense Manpower Data Center.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, *31*(2), 83-105.

Yen, W., & Fitzpatrick, A. (2006). Item response theory. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-154). Westport, CT: Praeger Publishing.

Zhang, B., & Stone, C. A. (2008). Evaluating item fit for multidimensional item response models. *Educational and Psychological Measurement, 68*(2), 181-196.

Zhang, J. (2012). Calibration of response data using MIRT models with simple and mixed structure. *Applied Psychological Measurement, 36*(5), 375-398.

CURRICULUM VITAE

LEANNE FREEMAN A.B.D., M.S., L.P.C.
414.467.1667, leannes4@uwm.edu

PROFESSIONAL PROFILE

Ambitious and motivated Doctoral Candidate in statistics and measurement seeking placement in educational testing. Brings 10 years of graduate study and experience in statistics along with a background in counseling and film production. Accomplished researcher in cutting edge topics. Experienced instructor for undergraduate, Master's and PhD levels. Natural communicator, highly organized, self-starter. Superior work ethic, exceptional teamwork and managerial skills.

EDUCATION

University of Wisconsin, Milwaukee — Milwaukee, WI
Doctoral candidate, Educational Psychology Statistics & Measurement
Minor, Political Science Statistics & Methodology
Projected graduation date: *December 2016*
Dissertation: Assessing model-data fit for compensatory and non-compensatory Multidimensional Item Response Theory models using Vuong and Clarke statistics

Mount Mary University — Milwaukee, WI
M.S., Community Counseling, 2007
Thesis: The life-management strategy of SOC: A validation study with U.S. undergraduate students

California State University, Long Beach — Long Beach, CA
B.A., Psychology with Statistics & Research emphasis, 2003

Educational Psychology *Completed doctoral coursework*

Structural Equation Modeling	Applied Multiple Regression and Correlation Analysis
Categorical Data Analysis	Advanced Experimental Design & Analysis
Multivariate Methods	Survey Analysis and Methodology
Instrument Development	Psychometric Theory and Practice
Item Response Theory	Cognition and Human Development courses

Political Science *Completed doctoral coursework*
Political Science Statistics I, II, and III
Advanced Political Science Methodology Bayesian Analysis

PROFESSIONAL EXPERIENCE

Teaching Assistant, Educational Psychology 724, Advanced Statistics
Fall 2014, University of Wisconsin - Milwaukee — Milwaukee, WI

- Course
  - Topics included hypothesis testing, power and effect size, ANOVA, multiple comparisons, factorial ANOVA, simple and multiple regression
  - Partnered with professor in presenting course material
  - Assisted in constructing assignments and exams
  - Responsible for grading and recording assignments and exams
  - Reviewed answers to assignments and exams with class on weekly basis
  - Tutored individual students and groups of students as needed

    - Primary point of contact for students
- Lab
    - Provided instruction in SAS and SPSS programming parallel to course content
    - Designed and taught all lab content
    - Facilitated deeper understanding of appropriate use of statistical procedures within varied research situations
    - Focused on interpretation and presentation of results

Instructor, Counseling 630, Statistics and Research Methods
Summer 2014, Mount Mary University Milwaukee, WI

- Statistics topics included hypothesis testing, sampling distributions, *t*-tests, ANOVA, simple regression, correlation, chi-square
- Research design topics included ethics and use of human subjects, basic research designs, evaluating and writing research papers (research questions and hypotheses, literature reviews, methods, results)
- Designed all in-class and online components
- Developed curriculum to adhere to state accreditation requirements
- Assessed student knowledge using on-line and in-class discussions and games, essays and short papers, projects, presentations, and exams
- Communicated expectations and created environment for student success
- Utilized variety of teaching techniques to engage students and convey concepts
- Facilitated research design projects that culminated into thesis proposals
- Instilled confidence in the understanding and use of basic statistics
- Promoted ability to critically analyze literature for evidence-based practices
- Held individual and group tutoring sessions on a regular basis and as-needed

Teaching Assistant, Education 325, Classroom Assessment
Spring 2014, University of Wisconsin - Milwaukee Milwaukee, WI

- Topics included reliability, validity, bias, creating educational goals and objects, assessment of students (traditional, observational, performance), test building, scoring, understanding standardized tests
- Assisted professor in presenting course material
- Graded all assignments, projects, and exams
- Facilitated exams

Intern
Summer 2012, University of Wisconsin – Madison, Department of Ophthalmology and Visual Sciences Madison, WI

- Analyzed data from The Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR)
- Downloaded and cleaned data, combined variables, and addressed missing data as needed
- Utilized appropriate and varied statistical processes to answer research questions
- Conferred with medical doctors regarding research questions, analyses, and results
- Wrote SAS programs from scratch for descriptive and inferential statistics
- Created plots, graphs, and tables as needed for analysis and presentation
- Presented results to colleagues and academics at quarterly summit

Tutoring
2006 - present Milwaukee, WI

- Tutored subjects including Item Response Theory, regression analysis (simple and multiple linear, categorical, logistic), analysis of variance and accompanying comparison tests, all basic statistics
- Worked with all ages and levels from high school through graduate school
- Collaborated with graduate students and practicing instructors in building and evaluating tests and surveys
- Consulted for theses and dissertations in formulating research question and hypothesis, creating appropriate design for optimum data collection and hypothesis testing, processing data and performing statistical tests, and analyzing and interpreting results

ADDITIONAL PROFESSIONAL EXPERIENCE

Counseling
2008-2011, Affiliated Clinical Services West Bend, WI
Individual, couple, group, and family counseling

- Created an encouraging and safe environment where clients could identify appropriate goals, explore alternatives, and alter behaviors and cognitions.
- Presenting issues included depression, anxiety, aggressive and defiant behavior, academic problems, marital dissonance, lack of interpersonal or coping skills, poor body image, and low self-efficacy.
- Created treatment plans and used suitable evidence-based theoretical approaches and techniques based on clients' stated goals.
- Evidence-based approaches included Cognitive Behavioral Therapy, Art and Music Therapies, Person Centered Therapy, and others as appropriate.
- Collaborated with professionals from varied mental health disciplines to both expand personal knowledge base and give clients optimal support.

2005-2008, Youth and Family Project West Bend, WI
Individual, group, and family counseling

- Group experience
    - Locations included Washington County Jail and Youth and Family Project office.
    - Facilitated group sessions for adults with children which encouraged communication, understanding, and knowledge of successful child/parent relationships.
    - Supervisor for Juvenile Detention Summer Program where responsibilities included planning daily activities, scheduling interns, and collaborating with jail personnel and local law enforcement.
- School-based experience
    - Locations included Badger Middle School, West Bend Alternative High School, West Bend East High School, and West Bend West High School.
    - Referral for at-risk teens in all locations as identified by school counselors.
    - Presenting issues included self-harm, aggressive behaviors, physical and emotional abuse, family conflict, interpersonal skills deficits, and academic challenges.
    - Collaborated with school counselors and families to facilitate academic and personal success.

Television Production
1994-2003, Independent contractor, various companies Los Angeles, CA
Executive Producer, Motion Graphics

- Responsible for all financial aspects of projects including bidding, actualization and invoicing.
- Supervised designers, animators, editors, and sound designers.
- Collaborated with creative directors on every aspect of project including layouts, color palettes, sound design and typography.
- Managed schedules and formats for public relations and advertising.
- Completed projects on time and under budget.
- Organized flow of projects and streamlined production.

Production, Live-action commercials and music videos

- Positions included Producer, Production Manager, and Production Coordinator.
- Worked with clients and directors to create desired product.
- Managed production from bid to delivery.
- Responsible for all financial aspects of the project.
- Negotiated, booked and supervised crew, equipment, locations, travel and talent.
- Calculated and processed union & non-union payroll.
- Actualized budgets, petty cash and per diems.

PROGRAM COMPETENCY
SPSS, SAS, R, Microsoft Office

LICENSE & AWARDS
Licensed Professional Counselor, awarded January 2010
Nominated Instructor of the Year by students, Mount Mary University, 2014-15

MEMBERSHIPS
American Educational Research Association
American Psychological Association, Division 5
American Statistical Association
National Council on Measurement in Education
Wisconsin Counseling Association